Stochastic Similarity Selections from Large Combinatorial Libraries

Victor S. Lobanov* and Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Suite 104, Exton, Pennsylvania 19341

Received August 31, 1999

A stochastic procedure for similarity searching in large virtual combinatorial libraries is presented. The method avoids explicit enumeration and calculation of descriptors for every virtual compound, yet provides an optimal or nearly optimal similarity selection in a reasonable time frame. It is based on the principle of probability sampling and the recognition that each reagent is represented in a combinatorial library by multiple products. The method proceeds in three stages. First, a small fraction of the products is selected at random and ranked according to their similarity against the query structure. The top-ranking compounds are then identified and deconvoluted into a list of "preferred" reagents. Finally, all the cross-products of these preferred reagents are enumerated in an exhaustive manner, and systematically compared to the target to obtain the final selection. This procedure has been applied to produce similarity selections from several virtual combinatorial libraries, and the dependency of the quality of the selections on several selection parameters has been analyzed.

INTRODUCTION

The explosive growth of combinatorial chemistry in recent years has been greeted as both a blessing and a curse. While it has solved the problem of throughput and has massively parallelized the traditionally slow and ineffective drug discovery process, it has created the need to deal with compound collections of truly staggering size. These include both physical collections of compounds that are synthesized using automated parallel synthesis, and virtual ones containing molecules that could *potentially* be synthesized by systematic application of established synthetic principles. The initial ambition to "make and test them all" gave way to a more pragmatic approach once it became evident that "all" was a number of immense proportions. For example, a simple diamine-based combinatorial library built from only commercially available reagents can include up to 10^{12} compounds, which is equivalent to ∼300 years of synthesis and testing at a rate of 10 million compounds per day.[1] The recognition of these practical limitations and the desire to use the available synthetic and screening resources in the most efficient way has generated interest in virtual chemistry and, in particular, methods for handling and analyzing large chemical libraries.[1,2]

A virtual library is essentially a computer representation of a collection of chemical compounds obtained through synthesis, acquisition, or retrieval. By representing chemicals in silico, one can apply cost-effective computational techniques to identify compounds with desired physicochemical properties,[3−6] or compounds that are diverse, or compounds similar to a given query structure.[1,7−9] By trimming the number of compounds being considered for physical synthesis and biological evaluation, computational screening can result in significant savings in both time and resources, and is now routinely employed in many pharmaceutical companies for lead discovery and optimization.

Virtual Combinatorial Libraries. Whereas a compound library generally refers to any collection of compounds assembled for a particular purpose (for example, a chemical inventory or a natural product collection), a combinatorial library represents a collection of compounds derived from the systematic application of a synthetic principle on a prescribed set of building blocks. These building blocks are grouped into lists of reagents that react in a similar fashion (e.g., A and B) to produce the final products constituting the library (C, Ai + Bj → Cij). Full combinatorial libraries encompass the products of every possible combination of the prescribed reagents, whereas sparse combinatorial libraries (also called sparse arrays) include systematic subsets of products derived by combining each Ai with a different subset of Bj's. Unless mentioned otherwise, the term combinatorial library will hereafter imply a full combinatorial library.

Once a synthetic protocol is designed, a virtual library can be created. In addition to providing the basis for computational screening, these libraries are also convenient for tracking and archiving purposes. The conceptual approach to generating virtual libraries is quite straightforward. The reaction transformations that convert the reagents into products is reduced to a set of substructure patterns which are mapped onto the reagents to identify reacting groups or atoms, and a list of instructions of how to modify the chemical graphs. These instructions include primitive actions such as removing an existing atom or bond, inserting a new bond between two atoms, or changing the order of a bond. For more complex reactions, the modifications may also include changing the formal charge or chirality of an atom. Complications arise from the fact that the substructure patterns have to be correctly defined so that they map only to those parts of a molecule that would indeed react under the prescribed conditions. For example, if one of the reagents is an amine, it can be defined as a nitrogen atom connected to a carbon and to at least one hydrogen atom (to account

* To whom correspondence should be addressed. E-mail: victor@3dp.com.

10.1021/ci990109u CCC: $19.00 © 2000 American Chemical Society
Published on Web 01/21/2000

for both primary and secondary amines). However, such a definition would also include amides, which are chemically different from amines. Hence, the definition of the amine pattern has to be extended to include further neighboring atoms or, alternatively, substructures which should be avoided have to be introduced. Furthermore, if primary amines are more reactive than secondary amines, primary and secondary amines have to be mapped separately and assigned different priorities. Additional issues that need to be addressed are removal of protecting or leaving groups (which may or may not be present) and handling of multiple possible products due to regio- or stereoisomerism, or the presence of multiple reactive functionalities within the same reagent.

By their very nature, combinatorial libraries can reach extremely large sizes even with a relatively modest number of building blocks, particularly if there is a large number of variation sites (a problem known as combinatorial explosion). For example, the above-mentioned diamine library can easily include 10^{12} compounds. Were it to be created, this library would contain 50 000 times more compounds than the world's cumulative chemical literature, and would require 10 terabytes of storage space at 100 bytes per structure. Finally, the enumeration of the entire library at a rate of 100 000 structures per second would take over 3 months. It is clear that when dealing with virtual libraries of this size explicit enumeration is not an option.[1]

Use of Virtual Combinatorial Libraries. The role of virtual libraries in drug discovery is to provide computational access to compounds that can be readily synthesized and tested for biological activity. The role of the computational tools is to identify which compounds from the library need to be tested to achieve the desired objective. In lead discovery, the objective is to explore different structural classes in order to identify "activity islands" in structure–activity space. Hence, the selected compounds have to be dissimilar from one another so that each compound provides unique and nonredundant information on the SAR landscape. Selecting compounds based on their dissimilarity or diversity has become very popular in recent years, and there have been an extensive number of publications which address this subject.[5,7,9–13] Once an initial lead (or leads) has been identified (that is, a compound that shows activity against the target and is structurally novel), the emphasis is shifted toward exploring more extensively the structure–activity space around that molecule. Typically, this is accomplished by selecting and screening compounds that are similar to the initial lead(s).[1,14] Finally, the accumulated qualitative and, if available, quantitative SAR information is used to optimize the initial leads into preclinical candidates through conventional medicinal chemistry techniques.

Besides diversity and similarity, other selection criteria can also be employed. Examples include selecting compounds having desired properties or property distributions as determined by a property prediction algorithm or a quantitative structure–activity model, or exhibiting an optimal fit to a biological receptor as determined by a biomolecular docking algorithm.[15–17] Substructure searching of virtual libraries appears to be less useful because their utility is different than that of conventional databases, and because it is superseded by similarity searching.[1]

Molecular Similarity. Similarity is one of the most subjective concepts in chemistry, and can be defined in a multitude of ways. Depending on the objectives, available tools, and other factors, compounds can be considered similar if they have similar numbers of atoms of the same types (constitutional similarity), similar numbers of bonds and rings of the same types and similar degree of branching (topological similarity), similar shape and surface characteristics (shape similarity), or similar electron density distribution (electrostatic similarity). Alternatively, similarity can be determined on the basis of the presence or absence of certain features such as a common substructure (substructural similarity), the relative position and orientation of important pharmacophoric groups (pharmacophore similarity), binding affinity as predicted by a receptor binding model (receptor affinity similarity), or the degree of conformational overlap with a known receptor binder (conformational similarity).

The precise numerical value that is used to describe the similarity between two compounds depends on the representation of these compounds, the weighting scheme used to scale different aspects of the representation, and the similarity coefficient used to compare these representations.[8] Often individual compounds are represented by a bit-string, such as a substructure key or a hashed fingerprint, where each bit or group of bits indicates the presence or absence of a particular structural feature. Alternatively, compounds can be represented by a vector of real numbers, each of which corresponds to a particular molecular descriptor. It has been suggested that in all cases the representation of the structures must comply with the "neighborhood principle" if it is to be useful in identifying biologically active molecules.[1,14] The neighborhood principle states that molecules with similar representations (i.e., molecules located within the same local region or "neighborhood" of the feature space) should have similar physicochemical properties. Recently, Patterson et al. have analyzed the neighborhood behavior of 11 sets of two- and three-dimensional molecular descriptors[14] following a validation study by Brown and Martin on the ability of these descriptors to cluster active compounds.[3] Descriptors which were found to exhibit "proper neighborhood behavior" included two-dimensional fingerprints, topomeric fields, and atom pairs. Finally, the degree of structural similarity between two compounds is quantified by means of a similarity coefficient, the most common of which are the Tanimoto coefficient for binary sets, and the Euclidean distance for real vectors. A thorough review of molecular similarity measures can be found in Willett et al.[8]

The Selection Problem. Searching a virtual library for compounds that are similar to a particular query structure or have a set of desired properties involves three steps for each compound: enumeration, calculation of descriptors, and evaluation of similarity or estimation of the property of interest. Due to the large number of products in many combinatorial libraries (particularly three- and four-component ones), just the enumeration part alone can take a few weeks of computational time, while the storage requirements can be prohibitive (vide supra). Since in these cases neither the generation nor the storage of fully enumerated libraries and their associated descriptors is feasible, we need methods that can identify the desired compounds without enumerating the entire library. One possible solution is to look at the far

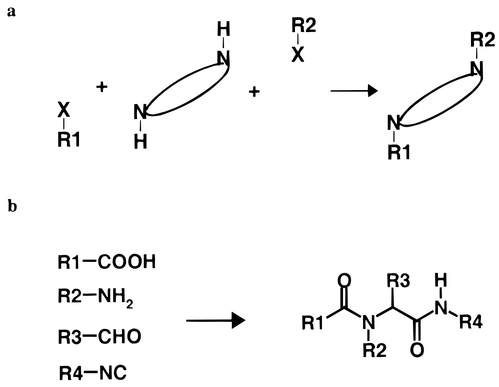


Figure 1. Summary of chemical transformations involved in generation of (a) diamine and (b) Ugi virtual libraries.

less numerous reagents instead of the products. The reagent-based approach is frequently used to maximize molecular diversity, and is based on the assumption that diverse reagents will lead to diverse products. However, it was recently shown that a selection based on the products themselves can be substantially more diverse, perhaps by as much as 35−50%.[13] When the selection criterion is similarity, the final products themselves must be considered, and the only proposed solution has been to use additive or otherwise "decomposable" descriptors. These are descriptors which, for combinatorial products, can be computed from the values of the corresponding descriptors of their constituent reagents.[1]

In this paper we present a novel stochastic approach for generating product-based similarity selections from large virtual libraries. This approach does not require enumeration and descriptor generation for the entire library, is not limited to additive or "decomposable" descriptors, and is able to provide an optimal or nearly optimal similarity selection in a reasonable time frame.

METHODS

Virtual Library Generation. A general, universally applicable program was developed for generating virtual libraries. The program takes as input lists of reagents supplied in SDF or SMILES format, and a reaction definition written in an extension of the scripting language Tcl. The use of a scripting language provides a powerful, human-readable, and convenient way of encoding chemical reactions. All chemically feasible transformations are supported, including multiple reactive functionalities, different stoichiometries, cleavage of protecting groups, and many others. The library is stored in a compact format without an explicit enumeration of the products. The computational requirements of the algorithm are minimal (even a billion-membered library can be generated in a few CPU seconds on a personal computer) and are determined not by the size of the library but by the number of reagents. Despite the implicit encoding, individual structures can be accessed at a rate of 1 000 000 per CPU second.[18]

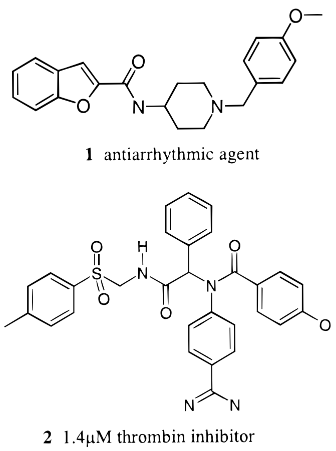
Two different combinatorial library designs were used in this work. The first was the diamine library that was described above, generated by combining a diamine core with a set of alkyl halides or acid chlorides[1] (Figure 1a). Although the physical synthesis of this library could prove problematic (the synthetic sequence involves selective protection of one of the amines and introduction of the first side chain, followed by deprotection and introduction of the second side chain), for the purpose of this study we assumed that one of the amino groups on the diamine core reacts with the first reagent, while the other reacts with the second reagent. A substructure search in the Available Chemicals Directory (ACD) yielded 1036 suitable diamines and 826 alkylating/acylating agents. These reagents were used to generate a virtual library containing over 706 million products (1036 × 826 × 826). Since the descriptors for the entire library could not be computed in a timely fashion, and since for validation purposes we needed to compare our results with conventional selections from a fully characterized library, a smaller 6.75 million membered library was produced by choosing 300 diamines and 150 alkylating/acylating agents at random. Hereafter, the term "diamine library" will refer to this smaller library, unless noted otherwise.

The second library was based on the Ugi reaction,[19] and involves an acid, an amine, an aldehyde, and an isonitrile (Figure 1b). A substructure search in the ACD yielded 1681 suitable acids, 594 suitable amines, 37 suitable aldehydes, and 17 suitable isonitriles. These reagents were used to build a virtual library containing over 628 million compounds (1681 × 594 × 37 × 17). Again, for validation purposes a smaller 6.29 million membered library was produced by choosing a random set of 100 acids and 100 amines. Hereafter, the term "Ugi library" will refer to this smaller library, unless noted otherwise.

Descriptor Generation. The evaluation of molecular similarity was based on a standard set of 117 topological descriptors computed using a C++ descriptor generation class from the DirectedDiversity API toolkit.[18] The descriptors included a well-established set of topological indices with a long, successful history in structure−activity correlation such as molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev−Trinajstis indices, and topological state indices.[20,21]

Similarity Evaluation. The calculated descriptors were normalized, and decorrelated using principal component analysis. The principal components which accounted for 99% of the total variance in the data (typically 25−30 principal components) were used to define the similarity space. Pairwise dissimilarity scores were calculated as Euclidean distances between the vectors associated with the respective compounds in the space defined by the selected principal components. A higher dissimilarity score indicates compounds that are less similar to each other, that is, more distant from each other in the principal component space. A candidate antiarrhythmic agent **1**[1] (Figure 2) was used as a query structure for the similarity searches in the diamine library because it can be formed by the diamine reaction sequence. Similarly, the 1.4 μM thrombin inhibitor **2**[19] originally devised by the Ugi reaction was used to derive similarity selections from the Ugi library.

As mentioned above, the concept of "neighborhood behavior" can be particularly useful in determining the ability of a particular descriptor set to discriminate between active and inactive molecules and assessing its relevance in similarity searching.[14] In order for a descriptor set to be a valid and useful measure of molecular similarity, a plot of pairwise similarity scores vs differences in biological activities for a set of related molecules should exhibit a characteristic trapezoidal distribution enhancement. Figure 3 dem-



1 antiarrhythmic agent

2 1.4μM thrombin inhibitor

Figure 2. Query structures used for deriving similarity selections from virtual libraries.[1,19]

onstrates the same characteristic behavior of our descriptor set and similarity evaluation procedure as applied to two recently published data sets.[22,23]

Compound Selection. Searching a combinatorial library for compounds that are similar to a particular query structure involves either direct or indirect comparison of the probe and the combinatorial products. When the size of the virtual library precludes enumeration, alternative searching techniques must be sought. Fortunately, in most cases the structures of the combinatorial products are determined to a large extent by the structures of their constituent reagents. Also, while the number of products can be huge, the number of reagents remains relatively small. This important feature of combinatorial libraries can be used to expedite the search for similar compounds if the following requirements are met: (i) all the members of a combinatorial library that are most similar to the probe are derived from a small subset of reagents and (ii) all the combinatorial products derived from any one of these reagents have, on average, higher similarity to the probe than the library as a whole.

The above requirements are, in fact, consistent with our chemical experience and intuition. Indeed, similar compounds tend to share common structural features, and in the case of combinatorial libraries, these structural features are partly contributed by the reagents that make up the individual products. On the other hand, compounds that are derived from the same reagents are likely to be more similar to each other. Consequently, if features of a particular building block are present in the probe, compounds that share that building block are likely to exhibit higher overall similarity to the probe than compounds that do not. Although it is possible to devise a combinatorial library and/or similarity measure that will not comply with the above requirements, in practice virtually all of them do. This is nicely illustrated in Figure 4, which shows the frequency of occurrence of each reagent in the top 100 compounds selected from the Ugi library based on their similarity to the thrombin inhibitor **2**. A similar discrimination of reagents into "preferred" (i.e., leading to compounds that are more similar to the query) and "indifferent" was also characteristic in the diamine library.

Profiling the reagents according to the dissimilarity score of all the products derived from these reagents exhibited the expected behavior as well. The similarity profile of a reagent was constructed by dividing the similarity scale into a number of bins and counting how many compounds derived from

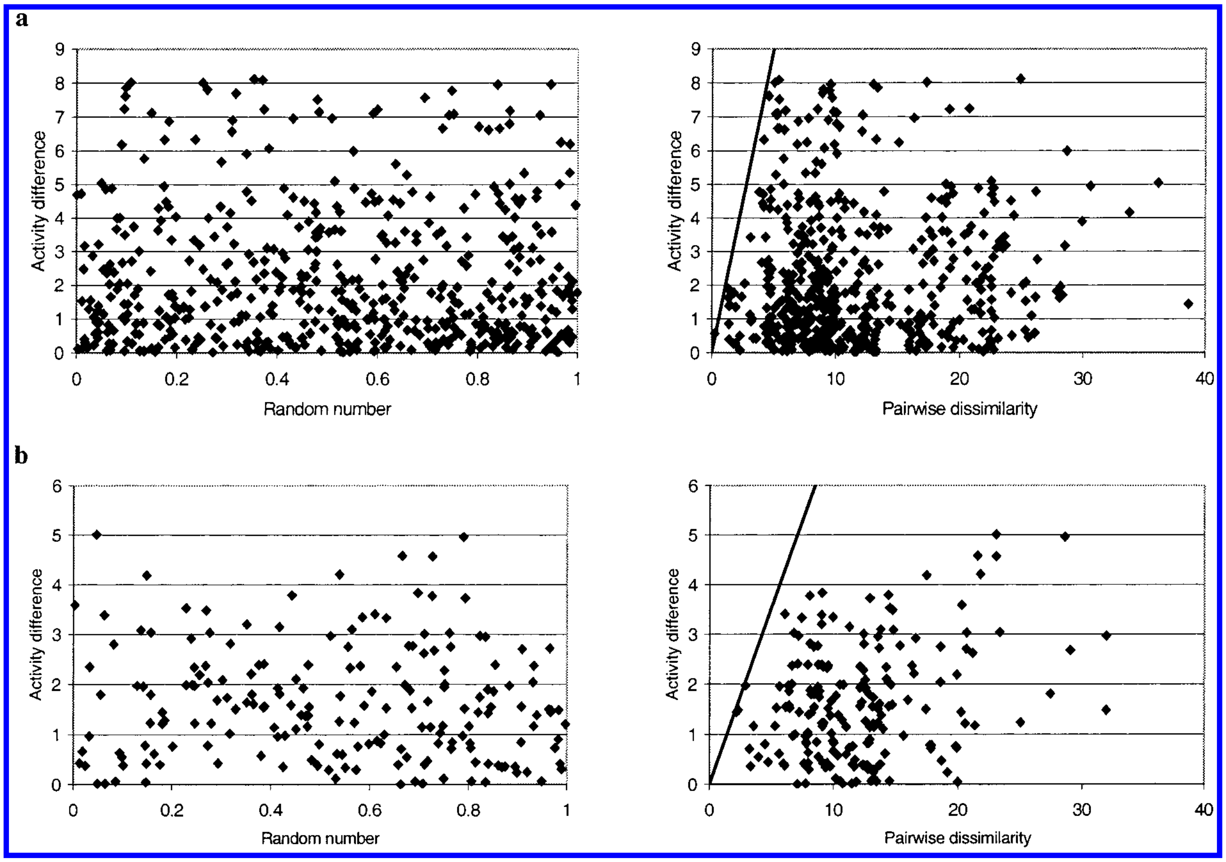


Figure 3. Neighborhood plots[14] comparing correlation between difference in biological activity and calculated pairwise dissimilarity and difference in activity and a random number for two (a and b) recently published data sets.[23,24]

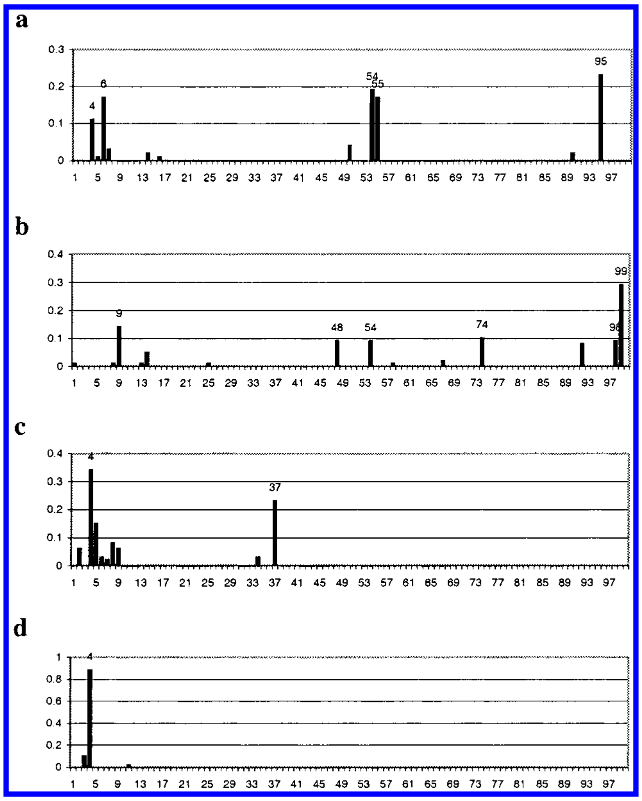


Figure 4. Frequency of occurrence of different (a) R1 reagents, (b) R2 reagents, (c) R3 reagents, and (d) R4 reagents in 100 most similar compounds selected from the entire 6.29M virtual Ugi library.

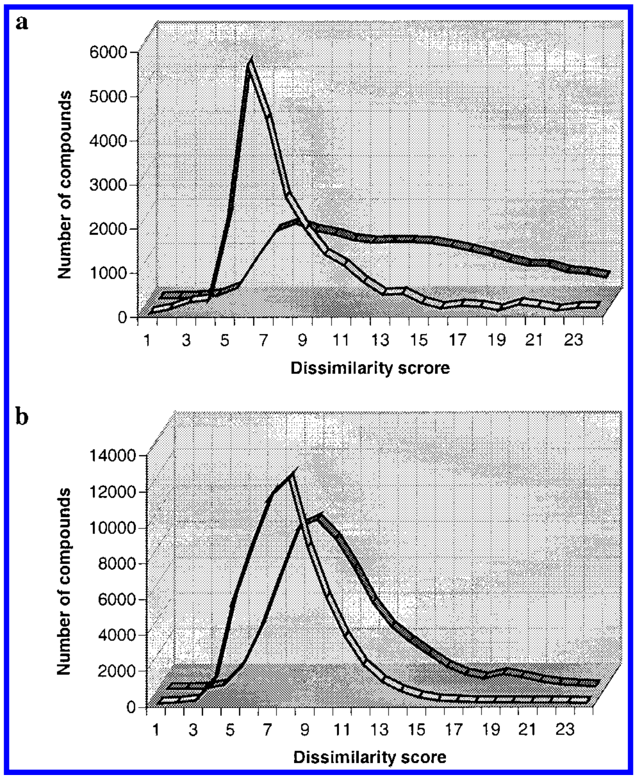


Figure 5. Typical similarity profiles of a "preferred" reagent (front profile) as compared to an average similarity profile (back profile) derived from (a) diamine and (b) Ugi virtual libraries.

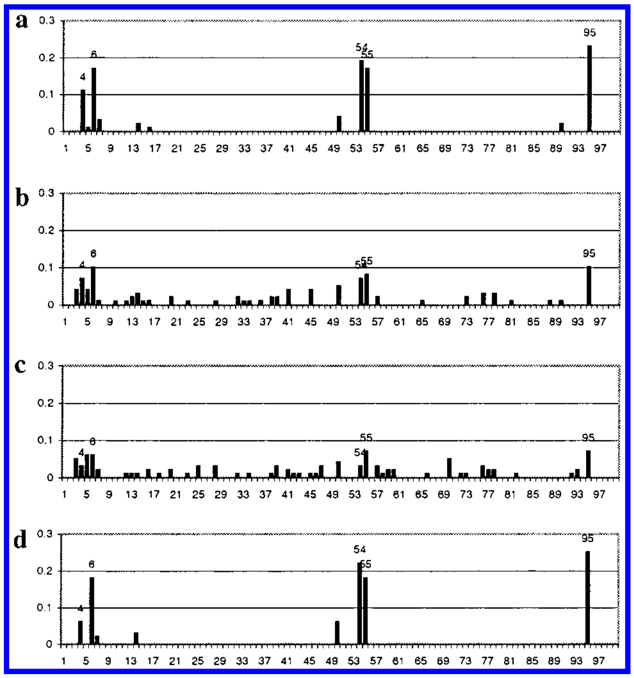


Figure 6. Frequency of occurrence of different R1 reagents in 100 most similar compounds selected (a) from the entire 6.29M Ugi library, (b) from a 500K random sample from the Ugi library, (c) from a 100K random sample from the Ugi library, and (d) by three independent 100K/50 runs. A description of the 100K/50 runs is given under Results and Discussion.

that reagent fall into each bin. Figure 5 illustrates the similarity profiles of two typical "preferred" reagents along with the average similarity profile, obtained by averaging the similarity profiles of all the reagents contributing to the same site of structural variation. Notably, the distribution of dissimilarity scores of combinatorial products derived from "preferred" reagents is always shifted toward lower values.

An important consequence of the above three properties of combinatorial libraries (namely, a large number of products being derived from a small number of reagents, the existence of "preferred" reagents or building blocks, and the characteristic shift in the similarity profiles of "preferred" reagents) is that it is possible to identify "preferred" building blocks by evaluating the similarities of a small, random subset of the products (Figure 6). The fact that each building block gives rise to a large number of products guarantees (in a statistical sense) that any sufficiently large random subset of products will "sample" each reagent many times. At the same time, the higher expected average similarity of products derived from "preferred" reagents will make these reagents "stand out", just as they do in the fully enumerated library. Once the "preferred" reagents have been identified, it is straightforward to find which combinations of them lead to the most similar compounds.

On the basis of the above, we propose the following stochastic procedure for fast similarity searching of large combinatorial libraries (Figure 7). First, a sufficiently large sample of products is selected at random from the library. The selected compounds (and only the selected compounds) are enumerated and characterized by calculating a prescribed set of molecular descriptors. The same descriptors are calculated for the query structure, and the pairwise similarities between the query and the enumerated combinatorial compounds are evaluated using a similarity measure of choice. The compounds are then sorted in descending order of their similarity to the probe, and the top-ranking compounds are deconvoluted into their building blocks. These

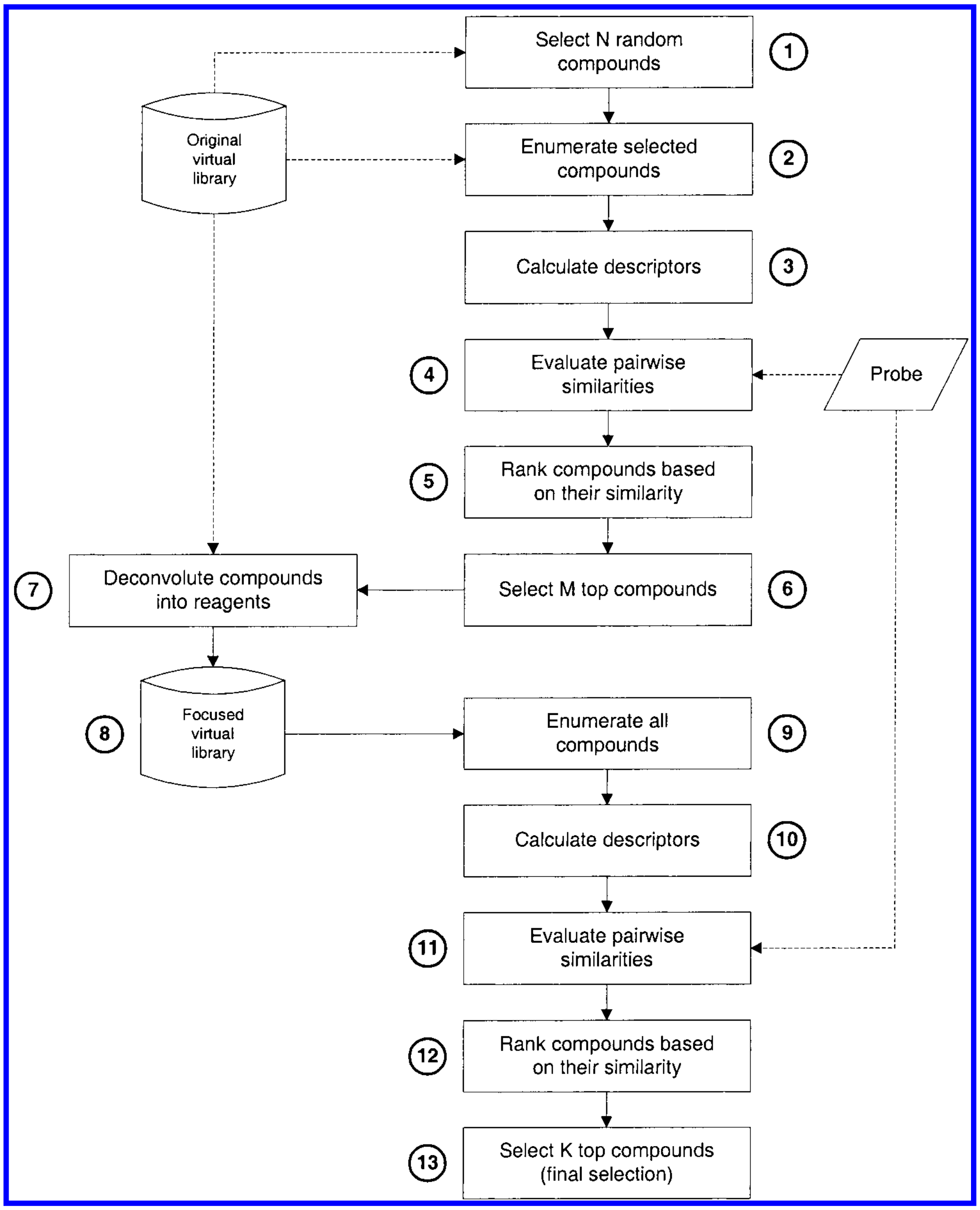

Select N random compounds
1
Original virtual library
Enumerate selected compounds
2
Calculate descriptors
3
Evaluate pairwise similarities
4
Probe
Rank compounds based on their similarity
5
Deconvolute compounds into reagents
7
Select M top compounds
6
Focused virtual library
8
Enumerate all compounds
9
Calculate descriptors
10
Evaluate pairwise similarities
11
Rank compounds based on their similarity
12
Select K top compounds (final selection)
13

Figure 7. Stochastic selection algorithm.

building blocks are subsequently combined into lists and are used to produce a smaller "focused" library which consists of all the cross-products of the selected reagents. All the compounds in that "focused" library are then enumerated, characterized by calculating descriptors, and compared to the query structure using the same similarity measure that was used in the screening of the original random set. Finally, the desired number of the highest ranking (most similar) compounds is extracted from the "focused" library based on their (dis)similarity scores. Because of its stochastic nature, the best results are obtained by repeating this procedure a few times and combining the results (cf. Figure 6a and 6d).

Software. The software is part of the DirectedDiversity suite[15−17] and is based on the Mt++ API toolkit developed at 3-Dimensional Pharmaceuticals, Inc.[18] All programs were written in C++ and were designed to run on all POSIX-compliant Unix and Windows platforms. Parallel execution on multiple CPUs is supported on both UNIX and Windows environment through the multithreading classes of Mt++.

RESULTS AND DISCUSSION

First, the entire 6.75 million member diamine library was enumerated, and the similarities of the virtual products to the query structure **1** were evaluated. Based on the calculated dissimilarity scores, a similarity profile of the diamine library was obtained by counting the number of compounds falling in each similarity bin (Figure 8a). According to the distribution, the majority of compounds had dissimilarity scores higher than 4. One hundred of the most similar compounds with the lowest dissimilarity score were selected and were used as a reference to compare all subsequent similarity selections drawn from the diamine library. These reference compounds represent the absolute best similarity selection of that size which can be obtained from the diamine library using the prescribed descriptors, similarity measure, and query structure.

Likewise, the entire 6.29 million member Ugi library was also enumerated and the similarities of the virtual products to the query structure **2** were evaluated. Although the Ugi library exhibited a more sharp similarity distribution (Figure

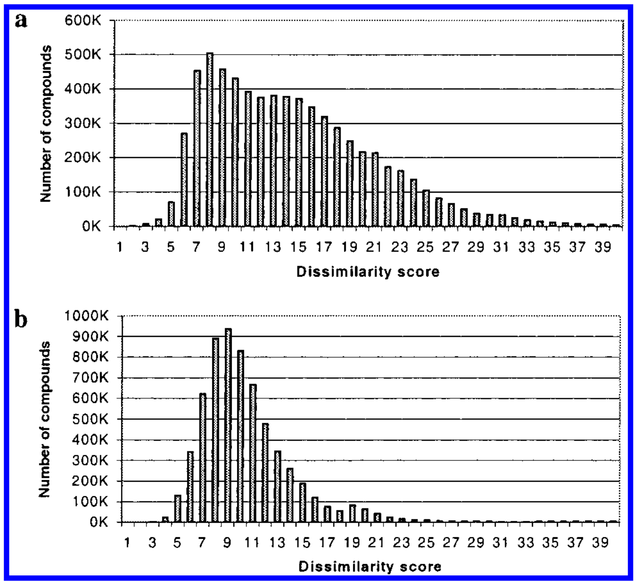


Figure 8. Similarity profiles of (a) diamine and (b) Ugi virtual libraries.

8b), the vast majority of the compounds in that library also had dissimilarity scores higher than 4. Again, the 100 most similar compounds were identified and were used as a reference to compare subsequent similarity selections from that library.

Next, the stochastic selection procedure outlined above was employed to select the 100 highest scoring compounds from the diamine library based on their similarity to the same query structure **1**. Initially 100 000 compounds were selected at random from the virtual library (step 1, Figure 7) and the 100 highest ranking compounds were used to produce lists of “preferred” reagents (step 6). The selection cycle was accordingly code-named 100K/100, and this naming scheme was used for all selections. The dissimilarity scores and IDs of the selected 100 compounds were compared with the dissimilarity scores and IDs of the reference selection derived from the fully enumerated library. On the basis of the average dissimilarity scores of the selected compounds (1.37 vs 1.30), the two selections were quite comparable. In fact, most of the compounds in the reference set were also found in the stochastic selection (Figure 9a). After repeating the stochastic procedure two more times with different random seeds and combining the results, the overlap with the reference selection rose to 96 out of 100 compounds. The same procedure was applied to the Ugi library, and an even better overlap with the reference selection was achieved (Figure 9b).

In three independent 100K/100 runs on the diamine data set, the 100 highest ranking compounds selected on step 6 were derived, on average, from 108 reagents, which produced a 50 000-membered focused library during step 8 (Table 1). Thus, even though only ~450 000 compounds (or 7% of the entire virtual library) were explicitly enumerated, described, and compared to the query structure over all three runs, 96% of the best possible hits were retrieved. In fact, the number of *unique* compounds screened is substantially lower due to the substantial overlap between the focused libraries generated in the three independent runs. For comparison, three independent screens of 200 000 random compounds selected from the diamine library retrieved a total of only nine (or

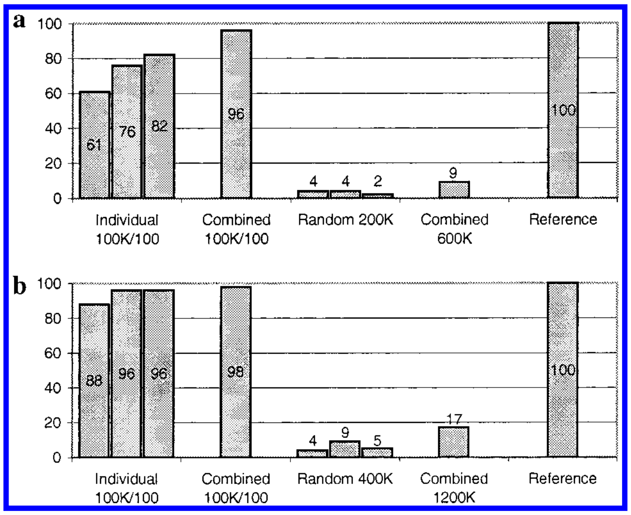


Figure 9. Overlap between stochastic selections and reference selections derived from (a) diamine and (b) Ugi virtual libraries.

three on average) out of the 100 most similar structures (Figure 9a). This is not surprising since 600 000 compounds constitute approximately 10% of the entire 6.75 million member diamine library.

In the case of the Ugi library, the 100 highest ranking compounds selected during step 6 produced an average of 99 “preferred” reagents. However, since this is a four-component library, the resulting focused libraries were larger—270 000 compounds on average. Thus, after three independent runs, a relatively higher proportion of the entire Ugi library was screened (18%), which probably explains the higher percentage (98%) of compounds recovered from the reference set (Table 1). Again, three random selections of 400 000 compounds each were able to find only 17 out of the 100 most similar structures (Figure 9b).

The efficiency of the proposed algorithm should be judged by two factors: (1) how good is the final selection and (2) how many virtual compounds were actually screened. These two criteria are naturally connected: for example, if all virtual compounds were screened, then the best possible selection would have been obtained. The goal is to find a nearly optimal set by comparing the smallest possible number of compounds, and thus complete the task in a reasonable time frame. The two parameters of the selection procedure that affect its outcome are the size of the initial random pool (step 1) and the number of the highest ranking compounds used to generate the focused library (step 6). To assess the effect of these parameters, a series of selections were carried out using several combinations of these parameters. For each combination, three independent runs were carried starting from a different random seed, and the results were combined and summarized in Table 1.

The number of highest ranking compounds used to determine the “preferred” reagents, which were used to produce the focused libraries, appears to have had the most significant effect on the quality of the final selection. This is not surprising, since a smaller number of compounds produces a smaller list of reagents which leads to a smaller focused library and, therefore, a smaller chance to retrieve the best hits. On the other hand, if that number is too large, much larger focused libraries are produced, and the execution speed of the algorithm is compromised (Figure 10). Thus, the optimal number must be determined on the basis of both

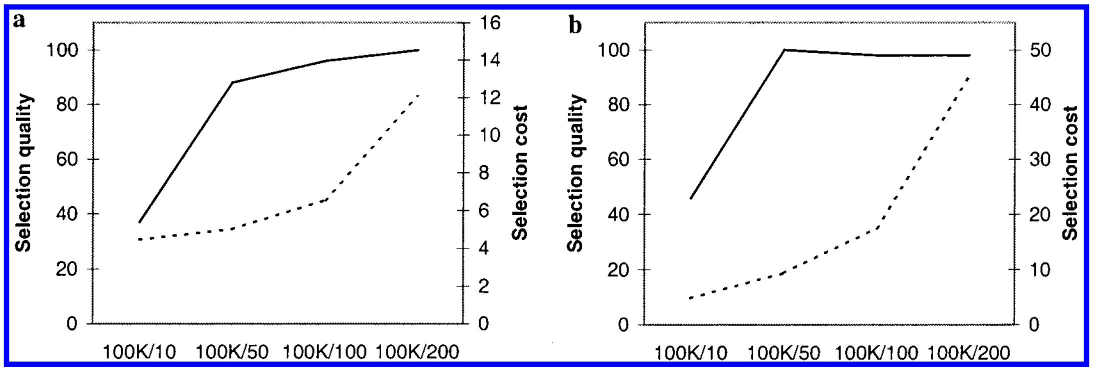


Figure 10. Effect of the number of top-ranked compounds chosen in step 6 on the quality and cost of the final selection from the (a) diamine and (b) Ugi virtual libraries. Selection quality (solid lines) is measured in the percent overlap with the corresponding reference selection. Selection cost (dotted lines) is the cumulative percent of the total number of virtual compounds evaluated.

Table 1. Summary of Similarity Selections

	size of random selection	no of top ranked compds selected	best similarity score	no. of reagents	size of focused library	no. of most similar compds selected	best similarity score	av similarity score	% of top 100 found	total compds screened	% of library screened
diamine library	n/a	100	0.7662	38	n/a	100	0.7662	1.30034	100	6 750 000	100
100K/10	100 000	10	1.0207	20	300	100	0.8585	1.48825	37	300 900	4
100K/50	100 000	50	1.2266	70	12 896	100	0.7662	1.31148	88	338 688	5
100K/100	100 000	100	1.0207	108	47 487	100	0.7662	1.30306	96	442 461	7
1K/100	1 000	100	2.1172	195	276 623	100	0.7662	1.32782	88	832 869	12
10K/100	10 000	100	1.6919	151	126 249	100	0.7662	1.30141	98	408 747	6
100K/100	100 000	100	1.0207	108	47 487	100	0.7662	1.30306	96	442 461	7
200K/100	200 000	100	1.4557	96	33 032	100	0.7662	1.30306	96	699 096	10
random 200K	200 000	100	1.1272	n/a	n/a	100	1.1272	1.82826	9	600 000	9
ugi library	n/a	100	0.0000	37	n/a	100	0.0000	1.47494	100	62 90 000	100
100K/10	100 000	10	0.7745	23	1 153	100	0.7745	1.78473	46	303 459	5
100K/50	100 000	50	1.6871	76	92 634	100	0.0000	1.47494	100	588 702	9
100K/100	100 000	100	1.4382	99	267 251	100	0.0000	1.47737	98	1 101 753	18
100K/200	100 000	200	1.1581	139	843 712	100	0.0000	1.47737	98	2 831 1366	45
1K/100	1000	100	2.5500	151	1 300 555	100	0.0000	1.47494	100	3 904 665	62
10K/100	10 000	100	1.3793	121	583 482	100	0.0000	1.48308	97	1 780 446	28
100K/100	100 000	100	1.4382	99	267 251	100	0.0000	1.47737	98	1 101 753	18
200K/100	200 000	100	0.77448	96	222 673	100	0.0000	1.47494	100	1 268 019	20
1K/50	1 000	50	2.50076	102	331 653	100	0.0000	1.52529	86	997 959	16
10K/50	10 000	50	1.22499	88	190 661	100	0.0000	1.49912	92	601 983	10
100K/50	100 000	50	1.6871	76	96 234	100	0.0000	1.47494	100	588 702	9
200K/50	200 000	50	1.1667	68	68 089	100	0.0000	1.48086	98	804 267	13
random 400K	400 000	100	0.0000	n/a	n/a	100	0.0000	2.18251	17	1 200 000	19

the quality and cost of the final selection. For the diamine library, 100 was the optimal number, whereas for the Ugi library just 50 compounds were sufficient to obtain the best selection. Undoubtedly, the optimal number of compounds to choose will vary from one virtual library to another, and will depend on the query structure as well. One can start with a small number and gradually increase it until the average dissimilarity score of the final selection riches a plateau (Figure 11).

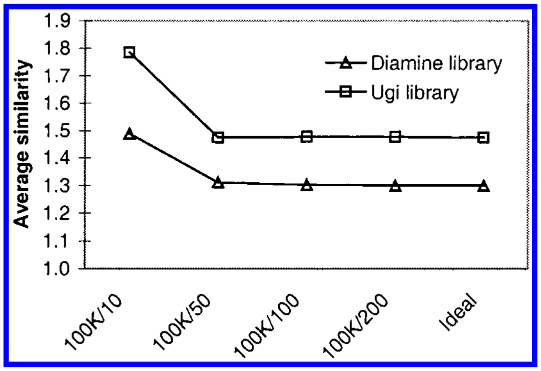


Figure 11. Average dissimilarity score of the final selection as a function of the number of the top-ranked compounds chosen in step 6.

Our experiments indicate that the size of the initial pool of compounds selected at random has a lesser effect on the quality of the final selection (Figure 12). However, if these initial pools are too small, they do not provide enough data to generate reliable statistics to determine the “preferred” reagents. If compounds are drawn from a virtual library at random, they will sample all the reagents with the same probability. For example, if only 1000 compounds are selected from the 300 × 150 × 150 diamine library, every R1 reagent will be present in only three compounds on average, while every R2 and R3 reagent will be present in only six compounds on average. In fact, the probability to “miss” at least one reagent from that library is almost 1 if only 1000 compounds are selected, whereas that probability is almost 0 if 10 000 compounds are selected. (The probability of missing at least one reagent by selecting M random compounds from a D-component virtual library $N_1 \times N_2 \times ... \times N_D$ can be roughly estimated by eq 1.) If the initial pool is too small, each of the highest ranking compounds can contribute completely different reagents, and the subsequent “focused” library can become too large and not focused at all. In this case, the similarity search degrades to

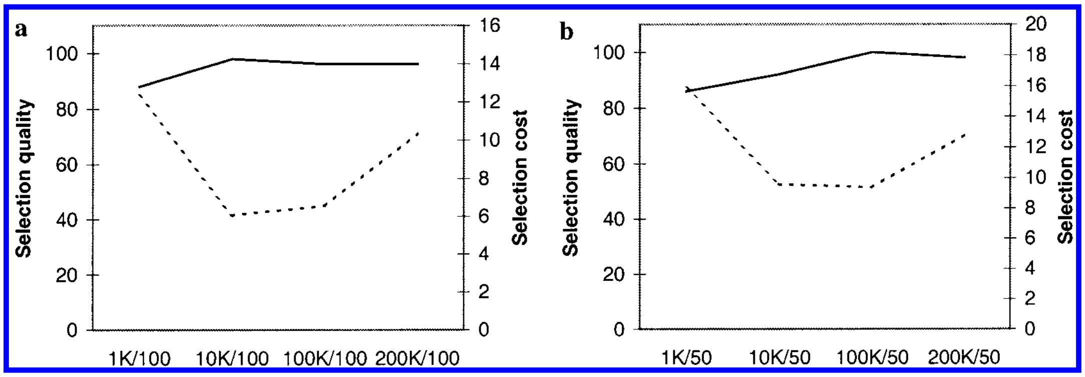
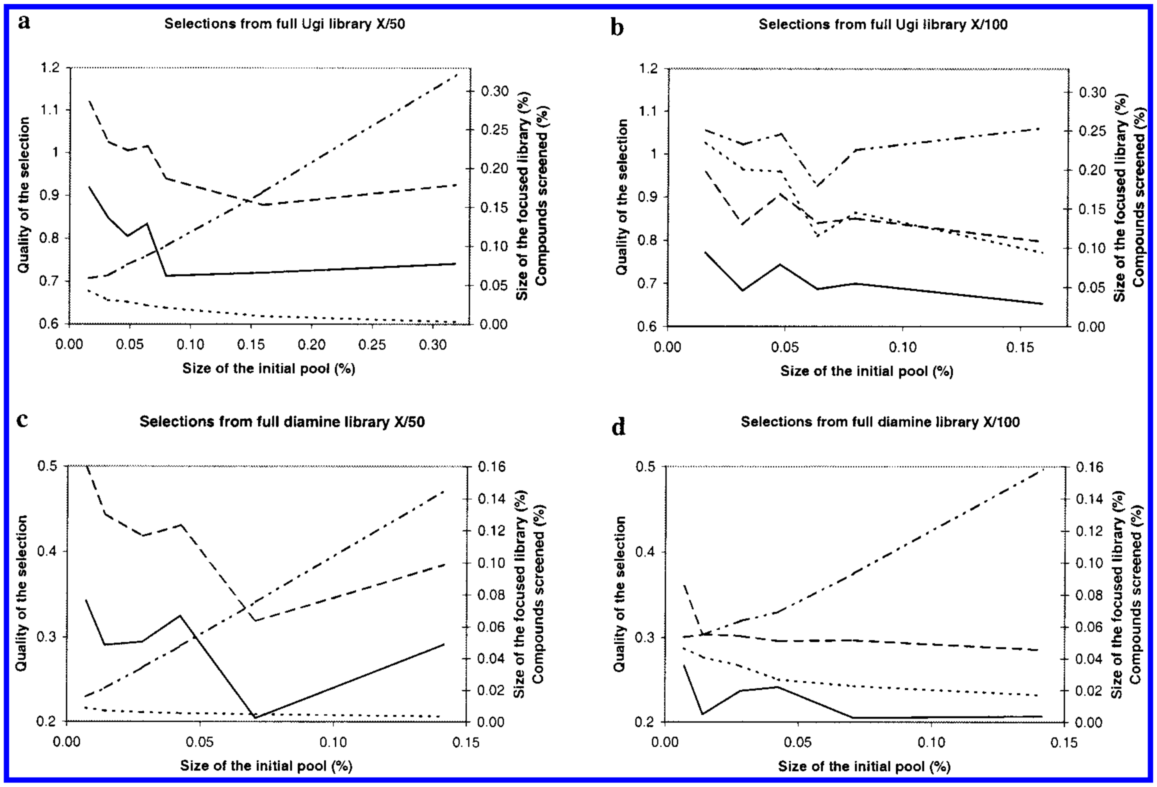


Figure 12. Effect of the size of the initial random pool of compounds on the quality and cost of the final selection from the (a) diamine and (b) Ugi virtual libraries. Selection quality (solid lines) is measured in the percent overlap with the corresponding reference selection. Selection cost (dotted lines) is the cumulative percent of the total number of virtual compounds evaluated.

$$p = \frac{\sum_{i}^{D}\left[N_i\left(\frac{\prod_{j}^{D}N_j - \prod_{k\neq i}^{D}N_k}{M}\right)\right]}{\left(\frac{\prod_{j}^{D}N_j}{M}\right)} \qquad (1)$$

a brute-force, random sampling approach, and becomes inefficient. On the other extreme, when the initial pool is too large, the number of "preferred" reagents and the resulting focused library decreases, but the cost of the initial screening increases. In our experience, approximately 0.1% of the compounds in a virtual library need to be tested in the initial stage to achieve a nearly perfect similarity selection and at the same time keep the experiment practical (Figure 13).

Since our algorithm is most useful when applied to massive libraries that are intractable by other means, we derived a series of selections from the full diamine and Ugi libraries containing 706 and 628 million compounds, respectively, using the same query structures **1** and **2**, and varying the same selection parameters (i.e., the size of the initial pool and the number of highest ranking compounds used to derive the focused library). As before, each combination of parameters was tested three times starting from a different random seed, and the results were averaged. The results are summarized in Figure 13. It is clear that excellent selections were obtained in both cases after screening on average less than 0.2% of the compounds in these libraries. Of course, in this


a Selections from full Ugi library X/50

b Selections from full Ugi library X/100

c Selections from full diamine library X/50

d Selections from full diamine library X/100

Figure 13. Dependence of the quality (average dissimilarity score) and cost (number of compounds evaluated) of the final selection on the size of the initial random pool of compounds for the full-size (a, b) diamine and (c, d) Ugi virtual libraries. Notably, the quality of the combined selections (solid lines) is substantially better than the average quality of the individual selections (dashed lines). With the increase of the size of the initial pool, cost of the selection is generally increasing (dash−dotted lines), whereas size of the focused library gets smaller (dotted lines).

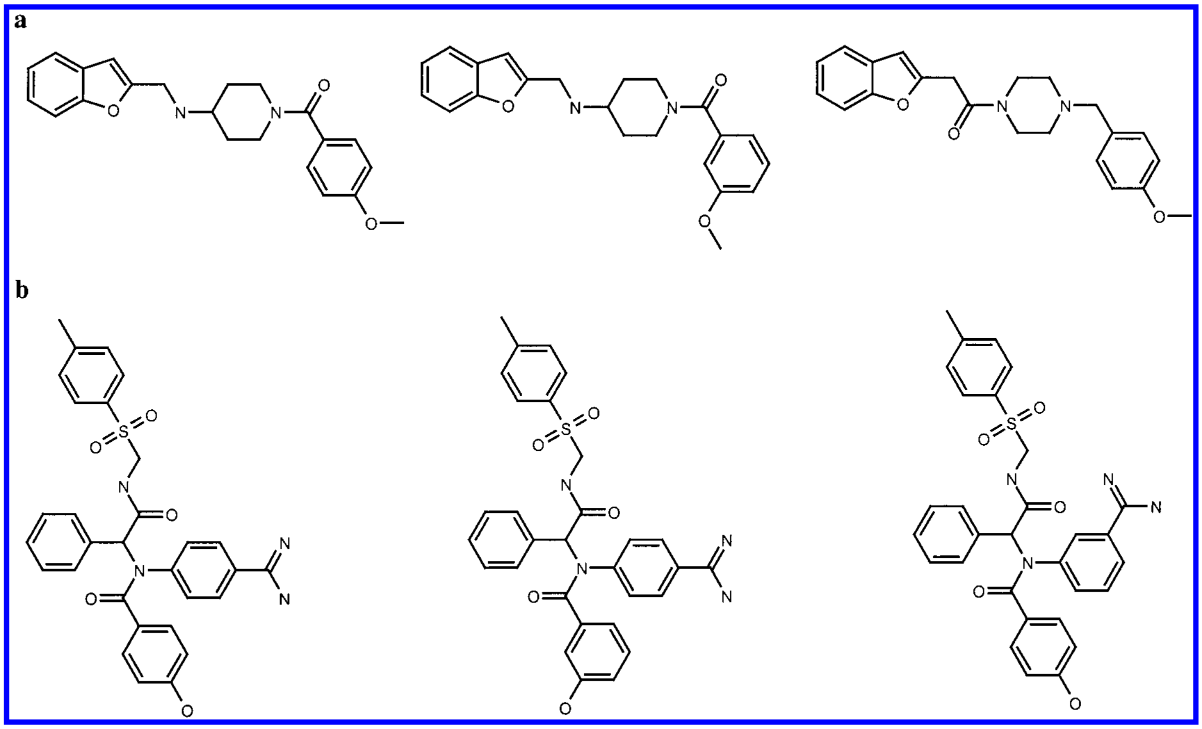


Figure 14. Samples of the most similar structures retrieved from the full (a) diamine and (b) Ugi virtual libraries.

case we do not know the best possible hits, but the average dissimilarity score of the selections derived from the full libraries is substantially lower than that of the corresponding selections from the smaller libraries (0.2 vs 1.3 and 0.7 vs 1.4 for the diamine and Ugi libraries, respectively). Figure 14 shows the structures of some of the most similar compounds found. The most important finding is that in every case the quality of a combined selection after three independent runs (solid lines in Figure 13) was substantially better than each individual selection (dashed lines). In fact, we found that better selections can be obtained by using a smaller initial pool and repeating the selection process three times than by running the selection once from a 3 times larger pool, in terms of both quality and speed.

These last selections also confirmed that for the diamine library choosing the 100 instead of 50 highest ranking compounds leads to a better selection, whereas for the Ugi library the improvement in quality is marginal and is outweighed by the increase in cost (dash−dotted lines). However, as we pointed out earlier, the two parameters affecting the outcome of the selection procedure are not entirely independent. When the initial pool is large and the number of highest ranking compounds selected is small, the resulting focused library is small. In fact, the larger the initial pool, the smaller the focused library (dotted lines). Since the number of highest ranking compounds chosen to generate the focused library remains constant, this means that these compounds contribute fewer reagents. A likely consequence of this is the possibility of missing some of the most similar compounds. Imagine, for example, that the single most similar compound in a virtual library is built from two reagents, the first of which, A, represents 70% of the product's structure and the second, B, the remaining 30%. Since the initial compounds are selected at random, it is very unlikely that the product AB will be picked. However, there will be several compounds containing only A (A-compounds) and several compounds containing only B (B-compounds). The larger the number of compounds screened, the more A- and B-compounds will be sampled. Because reagent A represents 70% of the product, the similarity of the A-compounds will be higher than that of the B-compounds. Therefore, if only a small number of the highest ranking compounds are chosen to generate the focused library, these will be exclusively A-compounds and the best compound AB will be never discovered. This effect can be seen in Figures 12a, 12b, 13a, and 13c, where the quality of the selection begins to decrease as the size of the initial pool becomes larger. Thus, a larger number of highest ranking compounds should be considered in such a case.

Performance. The most important quality of the algorithm presented herein is the ability to produce very good hit lists in a short period of time. When less than 1% of the virtual compounds need to be enumerated and characterized, the effective performance gain is 100-fold! Additional performance enhancements can be achieved by enumerating the compounds and calculating descriptors in parallel on multiple CPUs. For example, the enumeration and similarity evaluation of all 6.75 million compounds in the diamine library required 34 h on a dual processor 400 MHz Pentium II machine. The stochastic algorithm using an initial pool of 100 000 and a focused library derived from the 100 highest ranking compounds produced 88 of the 100 most similar compounds, and required only 30 min on the same system. A 1000K/100 selection from the 628 million membered Ugi library takes less than 2 h on a 6-processor R10,000 SGI. We believe that this performance represents a dramatic improvement over conventional methodologies, and allows these methods to be used in a routine fashion.

Future Directions. Similarity selections from virtual libraries are aimed at producing candidates for future synthesis and biological testing. To simplify the synthesis, combinatorial compounds are typically synthesized in an array format. Our stochastic procedure can be easily adapted to generate such arrays. After the "preferred" reagents have been identified and the focused library has been enumerated, a genetic algorithm or simulated annealing search engine[11] can be used to find an array that exhibits the lowest average dissimilarity score or highest percentage of hits. Furthermore, similarity is only one of many criteria that can be used to guide the selection algorithm. Examples including desired properties or property distributions, two- and three-dimensional QSAR predictions, and receptor complementarity will be presented elsewhere.[24]

CONCLUSIONS

A stochastic procedure for similarity searching of large combinatorial libraries was presented. This procedure avoids enumeration and descriptor calculation for every virtual compound, yet it provides an optimal or nearly optimal selection in a reasonable time frame. The algorithm was tested on two multimillion membered libraries and was shown to produce results that are comparable to those obtained from exhaustive enumeration. The algorithm is general and very fast, and can be used with any type of molecular representation and similarity measure.

REFERENCES AND NOTES

(1) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.

(2) Leland, B. A.; Christie, B. D.; Nourse J. G.; Grier, D. L.; Carhart, R. E.; Maffett, T.; Welford, S. M.; Smith, D. H. Managing the Combinatorial Explosion. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 62−70.

(3) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem*. **1997**, *40*, 2304−2313.

(4) Burden, F. R.; Winkler, D. A. New QSAR Methods Applied to Structure−Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci*. **1998**, *39*, 236−242.

(5) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem*. **1999**, *1*, 32−45.

(6) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(7) Clark, R. D.; OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1181−1188.

(8) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(9) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 169−177.

(10) Agrafiotis, D. K. On the Use of Information Theory for Assessing Molecular Diversity. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 576−580.

(11) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 841−851.

(12) Agrafiotis, D. K. Diversity of Chemical Libraries. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; Vol. 1, pp 742−761.

(13) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 731−740.

(14) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(15) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and Method of Automatically Generating Chemical Compounds with Desired Properties. United States Patent 5,463,564, 1995, and United States Patent 5,574,656, 1996.

(16) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System, Method and Computer Program for at Least Partially Automatically Generating Chemical Compounds Having Desired Properties. United States Patent 5,684,711, 1997.

(17) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System, Method and Computer Program for at Least Partially Automatically Generating Chemical Compounds with Desired Properties from a List of Potential Chemical Compounds to Synthesize. United States Patent 5,901,069, 1999.

(18) *The Mt Toolkit: An Object-Oriented C++ Class Library for Molecular Simulations*; 3-Dimensional Pharmaceuticals, Inc.: Exton, PA, 1994−1999.

(19) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization of the Biological Activity of Combinatorial Libraries by a Genetic Algorithm. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2280−2282.

(20) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Relations. In *Reviews of Computational Chemistry*; Boyd, D. B., Lipkowitz, K. B., Eds.; VCH Publishers: New York, 1991; Chapter 9, pp 367−422.

(21) Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys*. **1977**, *67*, 4517−4533.

(22) Lan, R.; Liu, Q.; Fan, P.; Lin, S.; Fernando, S. R.; McCallion, D.; Pertwee, R.; Makriyannis, A. Structure−Activity Relationships of Pyrazole Derivatives as Cannabinoid Receptor Antagonists. *J. Med. Chem*. **1999**, *42*, 769−776.

(23) Farutin, V.; Masterson, L.; Andricopulo, A. D.; Cheng, J.; Riley, B.; Hakimi, R.; Frazer, J. W.; Cordes, E. H. Structure−Activity Relationships for a Class of Inhibitors of Purine Nucleoside Phosphorylase. *J. Med. Chem*. **1999**, *42*, 2422−2431.

(24) Lobanov, V. S.; Agrafiotis, D. K. Manuscript in preparation.

CI990109U