Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets

Eric A. Jamois,* Moises Hassan, and Marvin Waldman

Molecular Simulations Inc., 9685 Scranton Road, San Diego, California 92121

Received March 4, 1999

With the current and ever-growing offering of reagents along with the vast palette of organic reactions, virtual libraries accessible to combinatorial chemists have dramatically increased in size. Yet, extracting representative subsets for experimentation is an essential step in the design of combinatorial libraries. There has been some controversy whether it is necessary to consider product properties, at some computational expense, or whether sufficiently representative sets can be identified from considerations of the reagent space alone. This study compares the efficiency of reagent-based selections and that of product-based combinatorial subsetting in the identification of representative library subsets. Quantitative estimates reported herein show that the advantage of working in product space is descriptor dependent. For some descriptors, product-based approaches provide a distinct advantage, whereas for others results from reactant pools offer comparable results. Hence the behavior of descriptors, in mapping diversity from reagent space to product space, should be investigated prior to embarking into lengthy product-based considerations. Several classes of descriptors are studied including two-dimensional fingerprints (ISIS and Daylight) and physicochemical descriptors.

INTRODUCTION

In the past few years, the efficient design of combinatorial libraries has become increasingly important for both lead identification and lead follow-up programs.[1−4] The vast palette of available reagents has had a large impact on the size of synthetically accessible libraries for both general screening and targeted applications. Extremely large libraries are often inaccessible and must be reduced to smaller subsets in order to accommodate current limitations of synthesis and screening equipment. Indeed, multistep reactions often involve common reactants for which large selections are available from commercial sources. The Available Chemicals Directory provides a large inventory of reagents from which an appropriate selection can be made. The problem therefore becomes the selection of the reagents that provide the best set of products. In this context, we define "best set of reagents" as the combination of building blocks that provides the set of products most representative of the entire library. In this endeavor, it is important to provide efficient methods for the subsetting of combinatorial libraries.

A number of techniques have been used toward the identification of library subsets.[5−17] A first category of techniques involves reagent-based selections, that is, selections involving reagent properties only. Such selections may be obtained using clustering techniques or dissimilarity-based methods. A second class of techniques involves the selection of diverse sets of products. Again, a range of clustering[16,17] and dissimilarity-based methods[6−15] has been used for selections at the product level. One of the limitations of dissimilarity-based methods applied to products is the lack of combinatorial constraints; thus the selections produced are synthetically inefficient. A third type of technique involves combinatorially constrained product selections. In this case, the combinatorial array may be maintained by the selection of reagents but the evaluation of diversity of the resulting subset is performed at the product level. Such a procedure using genetic algorithms has been previously described by Gillet et al.[18]

One major distinguishing feature between the techniques is the computational resource required to identify library subsets. Most reagent-based selection techniques require little computational resource due to the limited size of reagent lists. Also, the size of the problem is only additive with respect to the size of each reagent list. On the other hand, product-based techniques often require complex and time-consuming procedures due to the multiplicative nature of the problem. A reagent array of $50 \times 150 \times 200 \times 350$ for a four substituent system Rl × R2 × R3 × R4 would generate 525 million products. While the reagent lists can be handled separately by common analysis techniques, handling the full set of products would pose a serious challenge. The handling of diversity at the product level, by conventional techniques, requires full enumeration of the products, calculation of descriptors for the entire library, and efficient algorithms for deriving subsets. Dealing with problems of this complexity should be envisaged only in the perspective of substantially higher quality subsets than what could be obtained via simple reagent-based analysis. The analysis from Gillet et al.[18] suggests that higher quality subsets are indeed identified through product-based procedures. Our analysis weighs this benefit against the additional computational effort and investigates the possible descriptor dependency of the results obtained. We also investigate a number of methodologies for the derivation of near-optimum combinatorial subsets.

BACKGROUND

Combinatorial Optimization. Combinatorial optimization involves the identification of combinatorial subsets that

10.1021/ci990015k CCC: $19.00 © 2000 American Chemical Society
Published on Web 12/09/1999

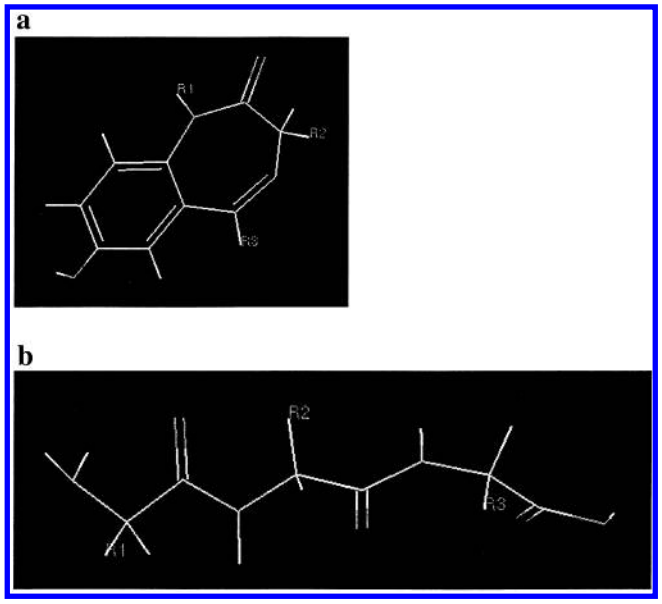

Figure 1.

adequately represent the entire library. Combinatorial subsets differ from simple product subsets in that they correspond to a strict array of reagents so that a given product cannot be obtained without generating the other cross products. A number of diversity metrics have been described by Hassan et al. to characterize noncombinatorial selections.[15] The measurement of diversity for combinatorial subsets poses different problems. Indeed, the combinatorial constraint usually imposes that a number of compounds be similar to each other in the resulting subset. This characteristic does not reflect the quality of the subset but simply results from its combinatorial nature. Although most distance-based diversity functions would not be well behaved under these conditions, pairwise dissimilarities have been used in combinatorial optimization procedures.[18] However, cell-based approaches can often provide faster measurements of space coverage.[19−21] A number of cell-based methods have been introduced herein to characterize combinatorial subsets.[22]

Descriptors. A number of descriptors have been investigated in their ability to characterize molecular diversity[16,17,23] ranging from MDL ISIS and Daylight fingerprints to topological indices. Investigations from Gillet et al.[18] revealed that, in the case of Daylight fingerprints, there are significant advantages to product-based approaches compared to simple reagent-based selections. As descriptors vary considerably in encoding molecular structure, there could be significant differences in the way diversity is inferred from reagent space to product space.

Measurement of Diversity and Space Coverage. Diversity and space coverage can be evaluated using a number of cell-based methods. These methods, implemented as diversity metrics,[22] evaluate how much of the space occupied by the complete library is filled by the subset. For example, the cell-based fraction metric attempts to select one compound from each cell in order to cover as many cells as possible. However, due to the combinatorial constraint, the objective to cover all occupied cells can seldom be achieved. The cell-based Chi^2 and entropy metrics attempt to level out the distribution to provide an even allocation of compounds to cells. The cell-based density metric attempts to select more than one compound from the most populated cells, in order to respect the level of occupancy of each cell. These metrics were used and compared as target functions in the combinatorial optimization process.

cell-based fraction: $F =$ cells occupied by subset/ number of occupied cells

cell-based Chi^2: $\chi^2 = \sum(N_i - N_{ave})^2$

cell-based entropy: $S = -\sum(N_i \text{ Log}(N_i))$

cell-based density: $D = -\sum(N_i \text{ Log}(N_i/M_i))$

where

$N_i =$ number of compounds in cell i for subset

$M_i =$ number of compounds in cell i for complete library

$N_{ave} =$ average number of compounds per cell expected for subset

$\sum =$ sum over cell occupied by subset

Subset Evaluation. Subset evaluation is required as an objective measure of the quality of the subset. The quality of the subset is defined as its ability to represent the entire library. Distance-based functions may be used toward this end to evaluate how much coverage of the entire library may be achieved by a given set of compounds.

METHODS

Test Libraries. We used the combinatorial library described by Ellman[24] as a first example (Figure la). The list of amino acids was reduced to eliminate enantiomeric reagents since the descriptors used in the subsequent analysis are not sensitive to stereochemistry. The resulting complete library consisted of a 16 × 18 × 20 array for R1 × R2 × R3 for a total of 5760 products. A library of tripeptides was used as a second example (Figure 1b). The complete library consisted of a 20 × 20 × 20 array for R1 × R2 × R3 for a total of 8000 products. In both cases, we investigated subsets of 672 compounds corresponding to a 7 × 8 × 12 array.

Descriptors. This analysis involves three sets of descriptors: ISIS keys (960 bits), Daylight fingerprints (1024 bits), and a set of 43 physicochemical descriptors. The set of 43 descriptors involved information content indices, structural descriptors, thermodynamic descriptors, and topological indices (Kier and Hall, Balaban, Wiener and Zagreb indices). These descriptors have been previously used in the characterization of molecules and fragments.[15,25]

Reagent-Based Selections. Our selection of reagents proceeds from a hierarchical cluster analysis (HCA) performed on each reagent list. The same set of descriptors was used for characterization of both reagents and products. In the case of physicochemical descriptors, principal component analysis (PCA) was first performed on the original set of descriptors. This procedure allowed for weighting of the principal components to compensate for possible correlations between descriptors.

The reagents were clipped prior to analysis and only retained the fragment portion attached to the scaffold. Clustering levels were chosen so as to obtain the desired number of compounds from each reagent list. The selections of reagents obtained were then directly mapped to the set of corresponding products since the complete library had already been enumerated. In this case, the selection of products obtained is combinatorial by nature since it proceeds directly from a selection of reagents. The combinatorial set of products was then submitted to the subset evaluation procedure described hereafter.

Combinatorial Optimization. Although distance-based measurements of diversity have been reported in combinatorial optimization procedures,[18] cell-based approaches can often provide faster measurements of space coverage.[19−21] However, a low-dimensionality space is required. A number of techniques are routinely used for dealing with the high dimensionality of the fingerprint space including multidimensional scaling (MDS).[26] We applied MDS to both MDL ISIS and Daylight fingerprints in order to provide a low-dimensional representation. However, in both cases, the first five MDS coordinates only explained 40−50% of the distance information contained in the native fingerprints.

An alternate approach involving clustering was adopted in order to provide a one-dimensional representation of the fingerprint space. Cluster analysis provides cluster membership information by assigning a cluster identification (ID) to each compound. Following on this idea, each cluster may be described as the equivalent of an occupied cell. In our case, the number of desired clusters is 672, corresponding to the number of desired compounds. Relocation clustering was identified as the method of choice since the number of clusters is known at the outset. Unfortunately, at the time of this study, our implementation of relocation clustering did not support pseudocentroids from fingerprints; HCA was therefore used instead.

Simple PCA was used when dealing with physicochemical descriptors. The first three principal components explained 84% of the original variance and were deemed sufficient to carry out the remainder of the analysis. The three principal components were used as the new coordinate system for combinatorial optimization and evaluation of the characterized subsets.

The combinatorial optimization uses a Monte Carlo procedure starting from a random $7 \times 8 \times 12$ array. The starting selection of reagents is then optimized in an annealing procedure involving 50 000 steps at $T = 1000$, 300, 100, 30, and 10 with a minimum of 5000 idle steps before the optimization is terminated. Given that the result potentially depends on the quality of the starting random set of reagents, 100 runs were performed corresponding to different seed values.

Random Combinatorial Selections. Random combinatorial selections were obtained and analyzed to provide a baseline in our analysis. A set of 100 random selections was obtained by the same process as described above but with 0 steps for optimization. The seeds used to obtain the random selections were the same as those used in the combinatorial optimization process. In this fashion, we could reproduce the original starting point and assess the improvement provided by each optimization trial. Indeed, it is possible that a better starting point provides better conditions for combinatorial optimization. Since the converse could also apply, it is important to assess the stability of the results obtained from combinatorial optimization.

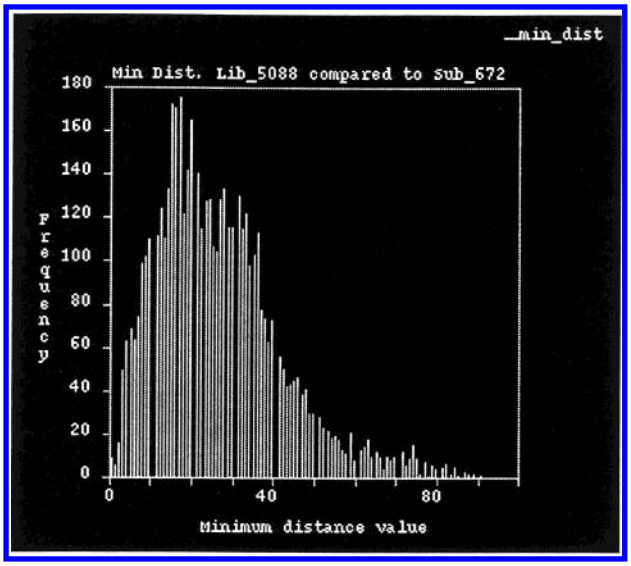

Figure 2.

Subset Evaluation. Evaluation of the identified library subsets is performed through a library comparison protocol. After n compounds have been selected as a combinatorial subset from a library of N compounds, we identify how well those $N - n$ compounds that were not selected can be represented by our selection. Thus, for every one of the $N - n$ compounds not selected, the closest compound in the selected subset is identified and the distance to this compound (Hamming distance or Euclidian distance) is computed. The distribution of distances can be plotted in a histogram (Figure 2).

A distribution histogram skewed toward very low distance values indicates that for every one of the unselected $N - n$ compounds a very close compound may be found in the subset. The selected subset is therefore highly representative of the entire library. Conversely, if the distance histogram is skewed toward large distance values, this indicates that the selected subset does not provide an adequate representation of the entire library. In some cases, the distance histogram may span a wide range of distances with a large proportion of occurrences for small distance values but also a number of occurrences for larger distance values. In this case, the selected subset may adequately represent a large proportion of the entire library but does not provide an adequate representation for some smaller set of compounds.

As an overall measurement of the quality of the subset, we calculated D_{mean} as the mean of minimum distances computed between each of the $N - n$ compounds in the remainder of the library and the subset. The lower the value of D_{mean}, the higher the representative quality of the selected subset.

Noncombinatorial Selections. Noncombinatorial selections were also obtained using a maximum dissimilarity algorithm (MaxMin)[15] driven by a Monte Carlo optimization similar to the procedure described earlier. In this case, more freedom is gained in the selection process through removal of the combinatorial constraint. This selection provided an approximate lower bound to D_{mean} obtained from combina-

Table 1. Values of D_{mean} Obtained in Subset Evaluation Experiments for Benzodiazepine Library

method	ISIS keys	daylight keys	physicochemical descriptors
random	39.70 (3.84)	31.96 (6.27)	0.293 (6.6 E−02)
reagent-based[a]	27.34	22.98	0.218
product-based: cell *F*	27.75 (0.92)	19.12 (0.62)	0.207 (8.8 E−03)
product-based: cell χ^2	34.00 (0.54)	19.98 (0.62)	0.200 (1.8 E−03)
product-based: cell *S*	32.84 (1.27)	19.09 (0.66)	0.202 (3.1 E−03)
product-based: cell D	25.94 (0.29)	16.84 (0.42)	0.226 (7.1 E−03)
noncombinatorial[a]	22.06	14.33	0.193

[a] Obtained from single runs. Other values are averages over 100 runs.

Table 2. Values of D_{mean} Obtained in Subset Evaluation Experiments for Tripeptide Library

method	ISIS keys	daylight keys	physicochemical descriptors
random	23.20 (6.37)	15.37 (4.21)	0.263 (4.3 E−02)
reagent-based[a]	17.07	11.72	0.214
product-based: cell *F*	13.77 (0.44)	7.93 (0.076)	0.180 (6.3 E−03)
product-based: cell χ^2	13.98 (0.38)	11.11 (1.51)	0.176 (5.9 E−03)
product-based: cell *S*	13.91 (0.44)	7.70 (0.00)	0.175 (4.2 E−03)
product-based: cell D	12.44(0.20)	8.80 (0.57)	0.169 (4.0 E−03)
noncombinatorial[a]	11.22	6.37	0.161

[a] Obtained from single runs. Other values are averages over 100 runs.

torial selections. This assumption is consistent with the arguments developed by Gillet et al.[18]

Given the CPU time requirements for this optimization procedure (1.5 h on a single CPU R10000 Silicon Graphics workstation), a single run was performed for each descriptor case.

RESULTS

Benzodiazepine Library. A summary of the results for the benzodiazepine library is provided in Table 1 with D_{mean} values corresponding to the different subset selection techniques. Numbers in parentheses indicate the standard deviation over selections for which 100 runs were performed.

Tripeptide Library. A summary of the results for the tripeptide library is provided in Table 2.

Benzodiazepine Library. Results for the benzodiazepine library are shown for ISIS keys (Figure 3), Daylight keys (Figure 4), and physicochemical descriptors (Figure 5). The graphs show the mean and range of D_{mean} attained for each method as well as standard deviation centered around the mean.

In the cases of ISIS keys, we observed limited benefits from the combinatorial optimization as the simple reagent-based selection performed quite well. Marginal improvements were obtained using the cell-density metric. In the case of topological descriptors, most optimization methods provided more representative subsets compared to reagent-based selection. However, the differences observed were still small. In the case of Daylight keys, more substantial differences were observed. This observation was consistent with the results reported by Gillet et al.[18] In this case, all combinatorial optimization methods provided significantly more representative subsets compared to the reagent-based selection.

We also compared the results from combinatorial optimization to the random subsets used as a starting point. The quality of the results from the optimization procedure showed little dependency on the starting point (seed) used, as indicated by the values of range and standard deviation for D_{mean}. This also confirmed the usefulness of the cell-based functions for selecting diverse and representative subsets.

Tripeptide Library. Results for the tripeptide library are shown for ISIS keys (Figure 6), Daylight keys (Figure 7), and physicochemical descriptors (Figure 8).

For all sets of descriptors investigated, the combinatorial optimization methods performed significantly better than simple reagent-based selections. This observation was also independent of the cell diversity metric used. As observed with the previous library, the combinatorial optimization methods provided very narrow ranges of D_{mean}, indicating a very stable behavior with regard to the starting point used.

DISCUSSION

Cluster Analysis. Cluster analysis provides valuable results that can be used for combinatorial optimization. As discussed earlier, MDS coordinates did not retain sufficient

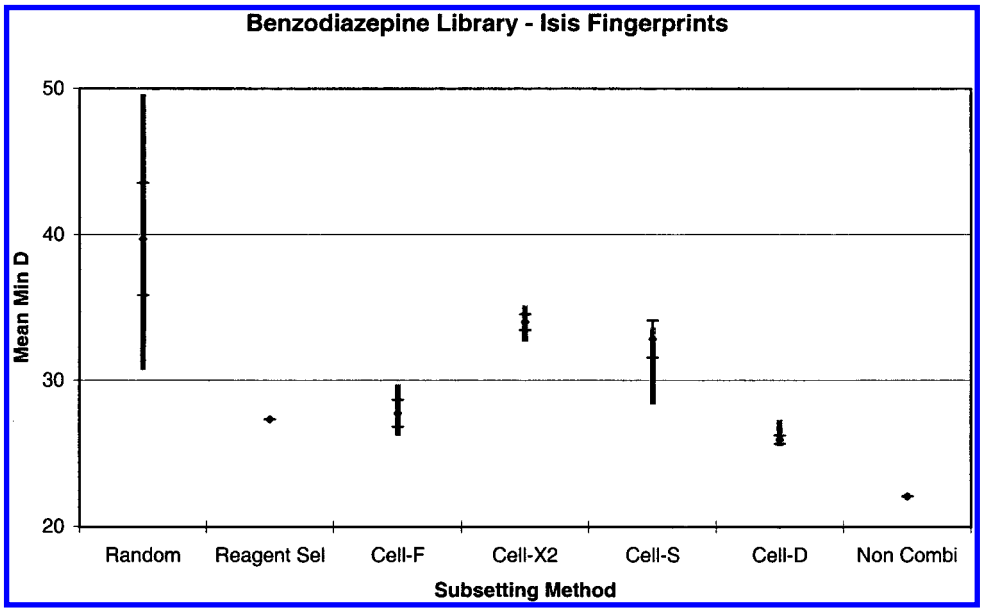

Figure 3.

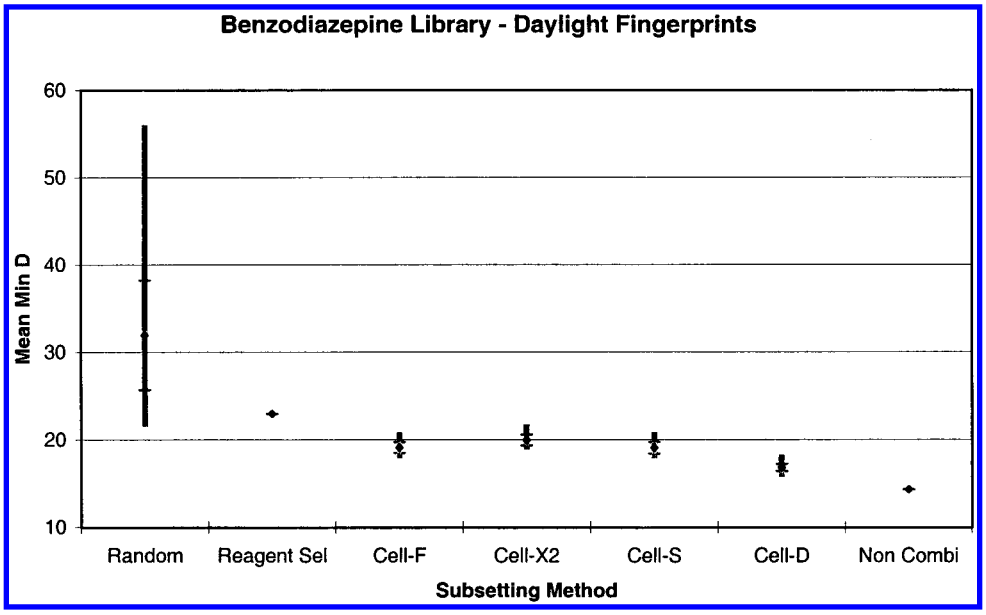

Figure 4.

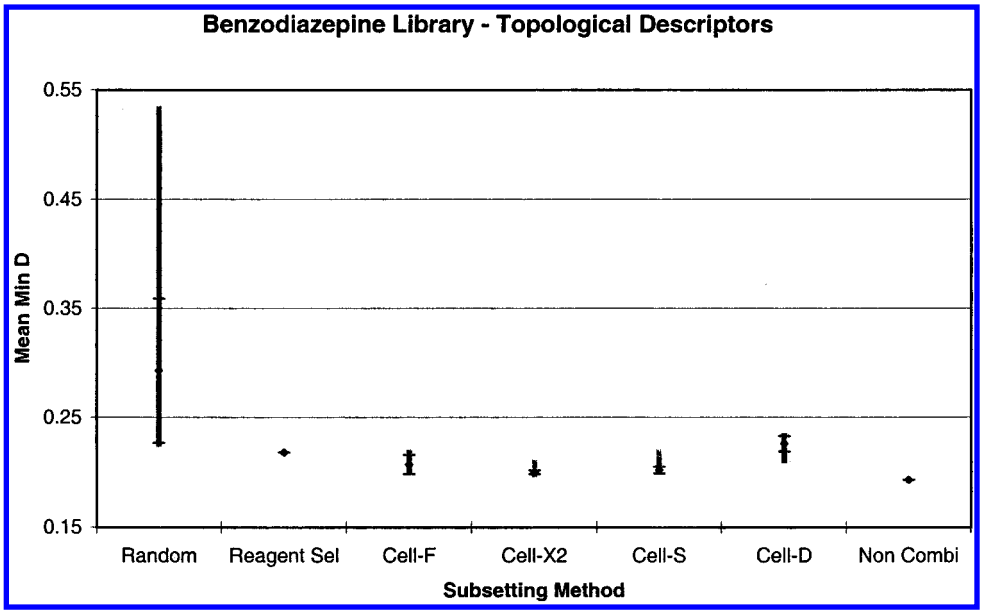

Figure 5.

distance information from fingerprints to be used as a lower dimensionality space. Cluster IDs assigned to compounds provided a useful one-dimensional representation from which combinatorial optimization could be performed. The superior results obtained with product-based combinatorial optimization compared to simple reagent selection attest to the usefulness of this technique. It should be noted that since the number of desired clusters is known at the outset, fast relocation clustering may be performed, even on large data sets.

Reagent-Based versus Product-Based. In our observations, the benefit of product-based combinatorial optimization compared to simple reagent selection depended very much on the descriptors used in the analysis and also on the nature of the library. In the case of the benzodiazepine library, using ISIS fingerprints, there was little or no improvement obtained from combinatorial optimization over a simple reagent selection. The choice of cell metric had only marginal effects on this outcome. With the selected set of physicochemical descriptors, the improvements were consistent but relatively small in regard to the added complexity of the product-based approach. On the other hand, when using Daylight fingerprints, the improvement was not only consistent but also substantial, highlighting the advantages of a product-based approach. This finding was consistent with the observations of Gillet et al.[18]

In the case of the tripeptide library, there were always significant advantages obtained from combinatorial optimization, regardless of the choice of descriptors. This behavior may be related to the degeneracy of the reagent lists. In the reagent-based approach, selection of R1, R2, and R3 are made independently of each other. For example, the selection of residues for R3 does not incorporate knowledge of the selections made for R1 and R2. This results in significant overlaps between R group lists. On the other hand, the combinatorial optimization provides simultaneous analysis and optimization of R1, R2, and R3 substituents.

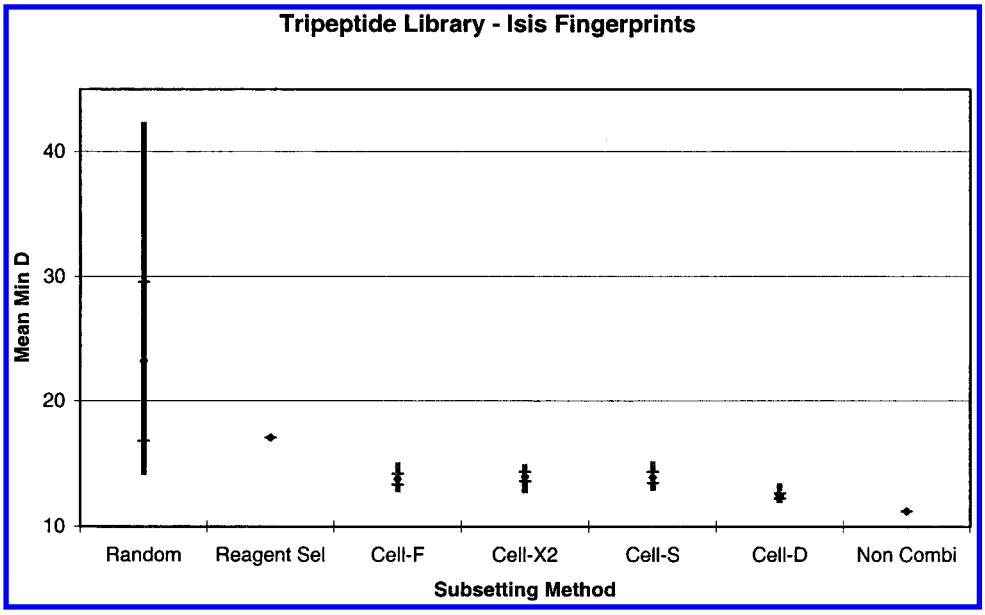

Figure 6.

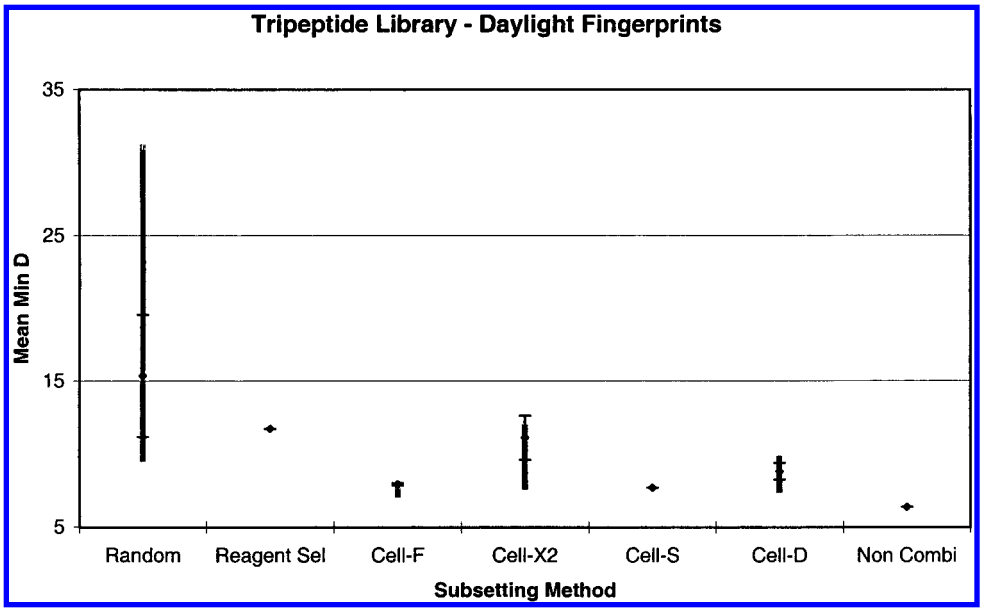

Figure 7.

Another key advantage of the product-based approach is the ability to provide additional constraints on the products such as "drug-like" characteristics based on whole molecule properties. These constraints can be managed in addition to the combinatorial constraints used in this study.

Descriptors. We observed substantial differences across classes of descriptors. Noteworthy is the difference between ISIS and Daylight fingerprints, which displayed very different behaviors in mapping diversity from reagent space to product space. It appears that, in the case of ISIS fingerprints, encoding of the products structural elements includes few extended paths and therefore does not reach into the core and across other R groups (Figure 9). Hence, information about each R group is possibly encoded independently of other R groups in the molecules. On the other hand, we suspect that Daylight fingerprints encode path information spanning from each R group into the core and possibly across other R groups in the molecules. In the case of a mixture of physicochemical descriptors, we expect that the behavior will depend on the nature of the descriptors used. In our case, we likely observed an average resulting from contributions from several individual behaviors. Descriptors such as low order connectivity indices span short paths and therefore should behave similarly to ISIS fingerprints. Conversely, high order connectivity indices span extended paths and therefore should behave similarly to Daylight fingerprints.

Combinatorial Optimization. The combinatorial optimization process attempts to identify a selection of reagents, which provides the best coverage of product space. In our case, the process optimizes a $7 \times 8 \times 12$ array from the complete array of $16 \times 18 \times 20$ (benzodiazepine library) or $20 \times 20 \times 20$ (tripeptide library) for R1 $\times$ R2 $\times$ R3. Even in case of the relatively simple benzodiazepine library, the total number of possible $7 \times 8 \times 12$ subsets is $C_{16}^{7} \times C_{18}^{8} \times C_{20}^{12} = 6.31 \times 10^{13}$, a formidable number. Since it would be impossible to systematically investigate every possible subset, we rely on the Monte Carlo procedure to provide a near-optimal solution. Related studies using genetic

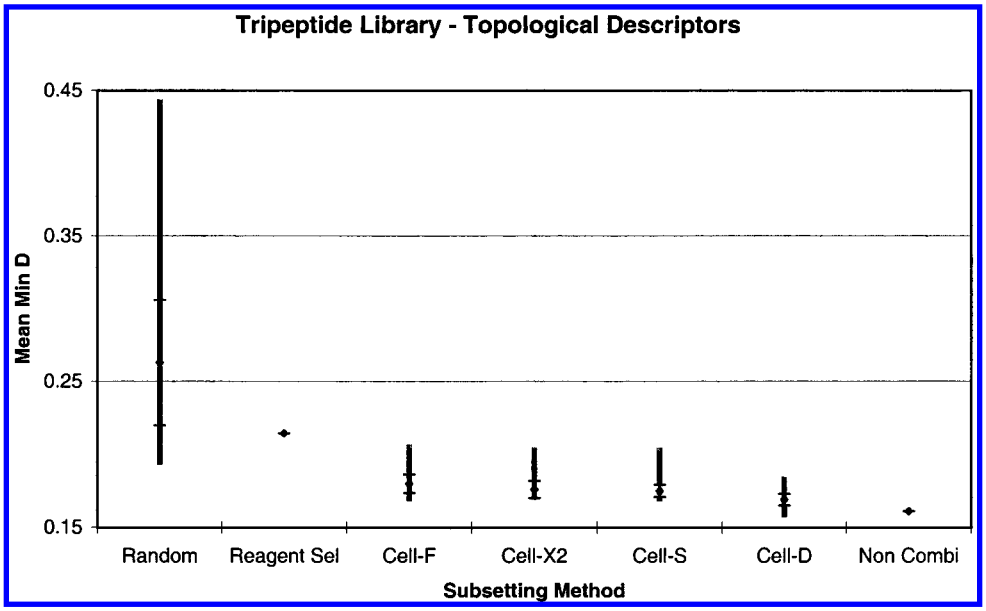

Figure 8.

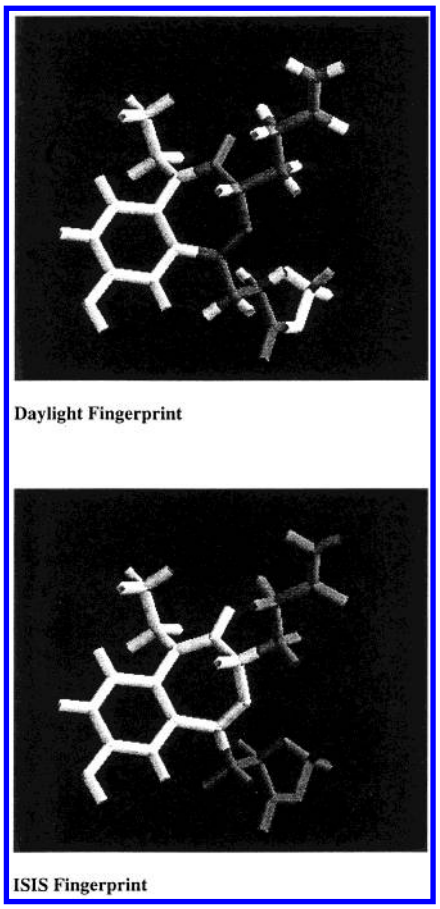

Figure 9.

algorithms suggest that the subsets obtained with such procedures are only slightly suboptimal.[18]

In order to quantify the stability of the solution provided by the combinatorial optimization, we compared the standard deviation in D_{mean} obtained in the Monte Carlo optimization procedure to the deviation in D_{mean} obtained in the random samples (Tables 1 and 2). We observed that the standard deviation from the optimization runs was considerably less than was obtained in the random runs. This result confirms the reliability of the optimization process with little dependency on the starting random selection.

Limitations. We recognize that, while considering a real combinatorial library, this analysis is purely theoretical. It would be interesting to compare the performance of subsets not purely from theoretical coverage considerations but also from experimental screening results.

Our investigations examined the dependence on several parameters but for only two libraries. It would be valuable to repeat similar experiments with different sizes and types of libraries. For example, libraries with different size common cores and different spacing of R groups could be investigated. We believe these factors may affect the differences observed between product-based and reagent-based approaches and their dependencies on the choice of descriptors.

Another factor that is likely to influence the mapping of diversity from reagent space to product space is the complexity of each reaction step involved in the elaboration of the combinatorial library. We envision that, if we introduce possible rearrangements depending on the nature of the reagents, then diversity will less obviously map from reagents to products. Under these conditions, there may be substantial advantages gained from product-based approaches.

CONCLUSION

Our results indicate that, in some cases, better subsets may be obtained through product-based combinatorial optimization compared to simple reagent-based considerations. The benefit gained from product-based approaches can highly depend on the type of descriptors used in the analysis. It may be ranked in the following order: Daylight fingerprints > physicochemical descriptors > ISIS fingerprints. We conclude that, for a chosen set of descriptors, a preliminary study should be performed to investigate the benefit of a product-based approach.

Other factors such as library size will influence the final approach since considerable effort may be necessary to enumerate large libraries and compute the chosen set of descriptors. The combinatorial optimization itself proved to

be very fast, requiring less than 1 min of CPU time for complete optimization using the annealing procedure described herein.[22]

In the case of three-dimensional descriptors such as those involving pharmacophore information, very little information about the products may be derived from individual consideration of the reagents. For this kind of property, product-based considerations may be the only possible approach.

ACKNOWLEDGMENT

The authors thank Dr. Robert Brown at MSI for helpful scientific discussions.

REFERENCES AND NOTES

(1) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(2) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(3) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J. A.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-Based Design and Combinatorial Chemistry Yield Low Nanomolar Inhibitor of Cathepsin D. *Chem. Biol.* **1997**, *4*, 297−307.

(4) *Combinatorial Chemistry;* Czarnik, A. W.; DeWitt, S. H., Eds.; American Chemical Society: Washington, DC, 1997.

(5) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity;* Wiley: New York, 1990.

(6) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Mod.* **1997**, *15*, 372−385.

(7) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(8) Lajiness, M. S. Dissimilarity-Based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.

(9) Lajiness, M. S. Molecular Similarity-Based Methods for Selecting Compounds for Screening. In *Computational Chemical Graph Theory;* Rouvray, D. H., Ed.; Nova Science: New York, 1990; pp 299−316.

(10) Bawden, D. Application of Two-Dimensional Chemical Similarity Measures to Database Analysis and Querying. In *Concepts and Applications of Molecular Similarity;* Johnson, M. A., Maggiora, G. M., Eds.; John Wiley and Sons: New York, 1990; pp 65−76.

(11) Lajiness, M. S. An Evaluation of the Performance of Dissimilarity Selection. In *Rational Approaches to the Design of Bioactive Compounds;* Silipo, C., Lajiness, M. S., Eds.; Elsevier Science: Amsterdam, 1991; pp 201−204.

(12) Marengo, E.; Todeschini, R. A New Algorithm for Optimal Distance-Based Experimental Design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37−44.

(13) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Structures from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(14) Holliday, J. D.; Willett, P. Definitions of "Dissimilarity" for Dissimilarity-Based Compound Selection. *J. Biomol. Screening* **1996**, *1*, 145−151.

(15) Hassan, M.; Bielawski. J. P.; Hempel, J. C.: Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Mol. Diversity* **1996**, *2*, 64−74.

(16) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(17) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(18) 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.

(19) Perlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf: Comput. Sci.* **1999**, *39*, 28−35.

(20) Mason, J. S.; McLay, I. M.; Lewis, R. A. *New Perspectives in Drug Design*; Dean, P. M., Nolles, G., Newton, C. G., Eds.; Academic Press: London, 1995; pp 225−253.

(21) Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.

(22) *Cerius*^{2}, Version 4.0; Molecular Simulations Inc.: 9685 Scranton Rd., San Diego, CA 92121.

(23) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(24) Ellman, J. A. Design, Synthesis and Evaluation of Small-Molecule Libraries. *Acc. Chem. Res.* **1996**, *29*, 132−143.

(25) Langer, T.; Hoffmann, R. D. New Principal Components Derived Parameters Describing Molecular Diversity of Heteroaromatic Residues. *Quant. Struct.-Act. Relat.* **1998**, *17*, 211−223.

(26) Everitt, B. S.; Dunn. G. *Applied Multivariate Data Analysis*; Oxford University Press: New York, NY, 1992.

CI990015K