*J. Braz. Chem. Soc.,* Vol. 13, No. 6, 788-794, 2002.
Printed in Brazil - ©2002 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

# Getting Past Diversity in Assessing Virtual Library Designs

***Robert D. Clark***

*Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 USA*

O planejamento de coleções combinatórias de compostos constitui, atualmente, um dos principais paradigmas da química medicinal. Muitos dos procedimentos publicados na literatura são "ótimos" por satisfazerem certos objetivos previamente definidos. Suas eficiências podem ser comparadas, mas em geral, torna-se muito difícil a comparação com coleções geradas por métodos diferentes ou mesmo por métodos que utilizem parametrizações diferentes. Isto é particularmente verdadeiro quando parâmetros outros que não os de diversidade molecular são importantes. Este artigo discute várias maneiras de comparar coleções de compostos, visual e numericamente.

A great deal of effort is currently going into the design of combinatorial libraries. Published approaches are generally "optimal" in that each best satisfies the target objective function it employs. Relative efficiencies can be compared in such cases, but it is often difficult to compare libraries generated by different methods or even by different parameterizations of the same method. This is particularly true once it is appreciated that attributes other than molecular diversity are important. This paper will discuss several ways in which library designs can be meaningfully compared to one another, visually as well as numerically.

**Keywords:** combinatorial library design, molecular diversity, representativeness, OptiSim, dissimilarity selection.

## Introduction

The incorporation of high-throughput screening (HTS) into the drug discovery and development process has prompted many pharmaceutical companies to shift from synthesis of individual compounds to combinatorial synthesis programs. This has led in turn to a broadening of the range of chemistry amenable to combinatorial approaches. One can now easily generate a virtual library composed mostly (if not entirely) of reasonably drug-like, synthetically accessible compounds the full realization of which would bankrupt any existing or conceivable pharmaceutical company many times over. This situation creates a pressing need for design tools to help chemists decide which particular products from such a virtual library should actually be made and tested. Many programs have been created for generating such sublibrary designs, with the most recent work focusing on choosing reagents so as to maximize some property of the specified products.[1] In many cases, the property being optimized is molecular diversity (substructural or pharmacophoric) among the products, though other objective functions have been used as well.[2]

* e-mail: bclark@tripos.com

In all cases, however, the intrinsic redundancy of combinatorial libraries guarantees that there will be many different solutions which are essentially equivalent with regard to the specific criterion being evaluated. Representativeness, in particular, is an important secondary consideration. Chemists need to be able to compare such alternative sublibraries in some general, detailed way. In addition, computational chemists need a meaningful way to evaluate the effectiveness of different design programs[3] and of reagent *versus* product-based design tools.[4] Here we use a series of sublibraries designed to be both representative and diverse to illustrate general analytical approaches based on Tanimoto similarities between substructural fingerprints. We use nearest neighbor similarity profiles and a recently developed variation on non-linear mapping for visualization to compare the various sublibraries.[5]

## Methods

### *Diversity analysis*

Molecular diversity was determined by comparing *UNITY*® substructural fingerprints[6] for the compounds in

question. These are bit vectors in which elements are set to 1 if particular substructures are present.[7] The similarity between fingerprints was evaluated in terms of the Tanimoto coefficient[8] $T$ as applied to bit set vectors **x** and **y**:

$$T(\mathbf{x},\mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|} = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x}| + |\mathbf{y}| - |\mathbf{x} \cap \mathbf{y}|} \quad (1)$$

where the bracketing vertical bars indicate cardinality. For diversity selection, the similarity between two sets is taken as the maximum similarity found between any member of one set and a member of the other, *i.e.*, the largest nearest neighbor similarity. This criterion underlies the maximal diversity selection algorithm[9] used in the *dbdiss* program distributed as part of the *Selector* module in *SYBYL*®.

A broader sense of the similarity of two (sub)sets can be obtained by examining the distribution of nearest neighbor similarities of one (sub)set with respect to the other. The nearest neighbor similarity profiles discussed here were obtained using the *dbcmpr* program, which is also part of the *Selector* module of *SYBYL*.

It is often enlightening to use *dbcmpr* to compare a set to itself – *i.e.*, to take the same set as target and reference. As discussed elsewhere,[10] it is particularly enlightening to do this when the set in question is a maximally diverse subset obtained using the algorithm embodied in *dbdiss*. The self-similarity profiles for such maximally diverse subsets can provide valuable insight into the structural scope of the data sets (here, sublibraries) from which they spring. The procedure is analogous to characterizing the area of a room by spreading a set number of dimes around it so as to maximize the distance (dissimilarity) between them. Having done so, the proximity (similarity) of the dimes characterizes the area (here, hypervolume) and shape of the room.

### *OptiSim selection*

Optimizable *k*-dissimilarity (OptiSim[11]) selection entails selection of the "best" candidate from each of a series of subsamples of size *k* drawn at random from the data set of interest.[12] Redundancy is prevented by checking each potential candidate against those items (here, compounds) selected in previous iterations; if it is too similar to any item already chosen, it is disqualified from further consideration. For most applications, a modified form of uniform random sampling without replacement is used, so that all potential candidates are considered before any candidate is reconsidered. The criterion used here to determine which candidate is "best" is structural diversity with respect to the compounds selected during previous iterations, with the first selection drawn at random or specified externally. To preclude structural redundancy,[13] candidates were excluded from subsamples if their fingerprints exhibited a Tanimoto similarity to those already selected greater than 0.90, corresponding to a "Tanimoto distance" ($d_T = 1 - T$) less than 0.10.

When applied to a large combinatorial library, the stochastic component of this simple strategy leads to selection of a set of compounds representative of the library as a whole. Choosing that candidate from each subsample which is least similar to those already selected, on the other hand, enhances the diversity of the selection set with respect to simple random selection. The balance between representativeness and diversity is set by the choice of *k*, with smaller subsample sizes favoring representativeness and larger sizes favoring diversity. Studies to date indicate that values of *k* in the range of 3 to 5 increase diversity without sacrificing much representativeness.[12]

### *Sublibrary block design*

Combinatorial sublibraries were created by applying an extension of OptiSim selection[5] in which successive reagent selections alternate between reagent classes. Consider, for example, two reagent sets **A** and **B** such that $A + B + X \rightarrow AXB$, where X is a common core or scaffold. Seed reagents $A_0$ and $B_0$ are selected at random. A subsample comprised of *k* candidates ($a_{11}, a_{12}, \ldots, a_{1k}$) chosen at random from **A** is then created, taking care that none of the products of reaction with $B_0$ (e.g., $a_{11}XB_0$) are too similar to $A_0XB_0$. The reagent leading to the product with the lowest Tanimoto similarity to $A_0XB_0$ is taken as the "best" candidate reagent; it becomes $A_1$. The design then pivots to consider reagents from **B**, with $b_{11}, b_{12}, \ldots, b_{1k}$ chosen at random, subject to the constraint that no product $A_iXb_{1j}$ is too similar to either $A_0XB_0$ or $A_1XB_0$. The best candidate $b_{1j}$ is then determined by identifying the one for which the similarity to the two products already selected is smallest. This candidate becomes $B_1$, the products $A_0XB_1$ and $A_1XB_1$ are added to the selection set, and the program proceeds to consider a new subsample of *k* candidate reagents from **A**.

In many cases, an unbalanced design is desired. If a larger reagent subset is specified for **B** than for **A**, pivoting stops once the quota for As has been fulfilled and subsamples of reagents are drawn from **B** until the block is completed. A new block is then initiated by drawing *k* products at random from the parent library and comparing them against all products included in the first block, and a new block is grown. The pattern of product selection produced by application of this method is illustrated in Figure 1. The designs described here were produced using

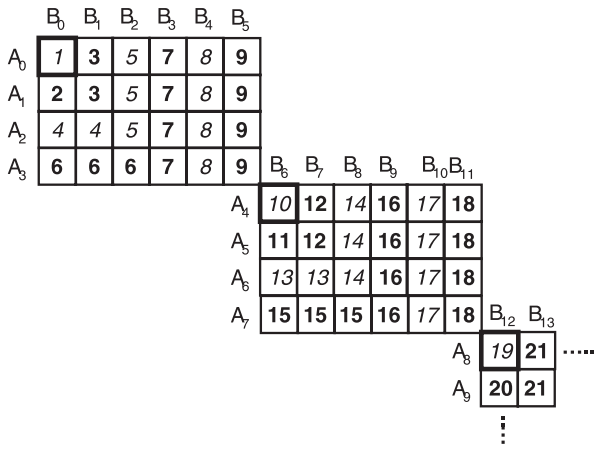



**Figure 1.** Schematic illustration of the order of selection for the first 52 products generated by the OptiSim multi-block 4x6 combinatorial matrix design program. The inset numbers indicate the iteration at which the corresponding product ($A_iXB_j$) was selected. Italics and boldface print set off products from each level of reagent selection ($A_1$ and $B_1$, $A_2$ and $B_2$, etc.). See text for details.

a prototypical implementation of the method written in SYBYL programming language (SPL).

It bears noting that filters other than simple redundancy – *e.g.*, acceptability of expected physical properties - can readily be put in place for determining the eligibility of candidate reagents for the subsample. Similarly, the "best" candidate in each subsample need not be determined by structural diversity, as it is here; similarity to a lead compound or incremental goodness of fit of the selected population to some target profile can be substituted. It should also be noted, however, that applying the diversity criterion to a *series* of subsamples rather than to the library as a whole serves to shift the properties of the sublibraries obtained away from simple diversity.

*Non-linear mapping with horizon (NLM-H)*

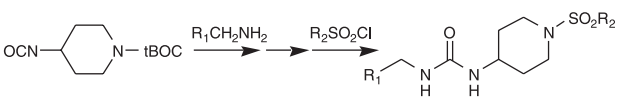
*UNITY* substructural fingerprints are made up of 988 binary elements. It is impossible for a human being to directly perceive relationships in such a high-dimensional space, and the Cartesian space to which we are accustomed is not the most appropriate one for making such comparisons anyway. As noted above, the Tanimoto similarity coefficient is better suited for this purpose, but it can only be directly applied to pairwise comparisons. Hence a tool is needed which can project most of the relevant information contained in the 988-dimensional "fingerprint space" down into two or three dimensions without unduly distorting important underlying Tanimoto relationships.

One way to accomplish this is by using principal components analysis (*PCA*) to get initial coordinates, and then using non-linear mapping (*NLM*)[14] to relieve distortions created by that projection. Behavior in such projections is dominated by long-range relationships, however, whereas it is *local* similarities that contain the most important information in fingerprint space; differences between low similarities tend to be meaningless.[15] Worse, long range relationships in this space are intrinsically very high dimensional, leading to large residual distortions in the projections obtained. Local relationships, on the other hand, tend to be of relatively low dimensionality, because the space is typically quite sparse. The best strategy in such a case is to modify the stress function so that any Tanimoto distance beyond a particular "horizon" contributes nothing to the aggregate stress for the projection unless it is projected to fall inside the horizon.[5] This *NLM-H* procedure is equivalent to cutting the space through unoccupied areas of the fingerprint space and unfolding it into two or three dimensions.

For the work described here, the *NLM-H* horizon *h* was set to a Tanimoto similarity of 0.70 ( *i.e.*, $d_T = 0.30$). Isolated compounds – those for which all other compounds in the set under consideration fall beyond the horizon – are placed at the border of the plot in a "hedge" of singletons. The plots shown here were obtained using a prototype version of *SARNavigator.*[16]

## Results and Discussion

A virtual library composed of 4-ureidopiperidines was created using the Selector and Legion modules of the *SYBYL Molecular Diversity Manager*.[17] This entailed using *UNITY*[6] to search for commercially available[18] reagent candidates bearing ten or fewer rotatable bonds, filtering out those containing undesirable substructures and exhibiting physical properties (molecular weight or volume, estimated hydrophobicity, etc.) too different from those exhibited by known drugs; details have been presented elsewhere.[5] The 308 primary amines and 154 sulfonyl chlorides so identified were then "reacted" *in silico* to give a virtual library of 47,432 potential products:




Three sublibraries comprised of 200 members each were drawn from this parent library. One was generated by applying "classical" OptiSim selection. The products included in this *cherry-picked* design come from entirely

unconnected reactions, so 200 different amines and 200 different sulfonyl chlorides would be required were the sublibrary to be realized through actual synthesis. A second, *single block* sublibrary was generated by selecting 20 amines and 10 sulfonyl chlorides as described above, with all possible cross-products included in the sublibrary. A third, *four block* sublibrary was generated wherein each block was made up of products resulting from all possible combinations of 10 amines and 5 sulfonyl chlorides. The library design program selected 20 distinct sulfonyl chlorides for this design but only called for 32 amines. Fewer than 40 amines were required because candidate reagents are selected with replacement. Hence the same amine can – and does – show up in different blocks.

All three designs were created using a subsample size *k* of 5, the same initial product, and the same random number seed. This produced a product distribution similar to that illustrated schematically on a smaller (48 compound) scale in Figure 2. Note that all three designs have some products in common, including the "seed" compound, and that the first block of products from the four block sublibrary is, by design, wholly included in the single block sublibrary.

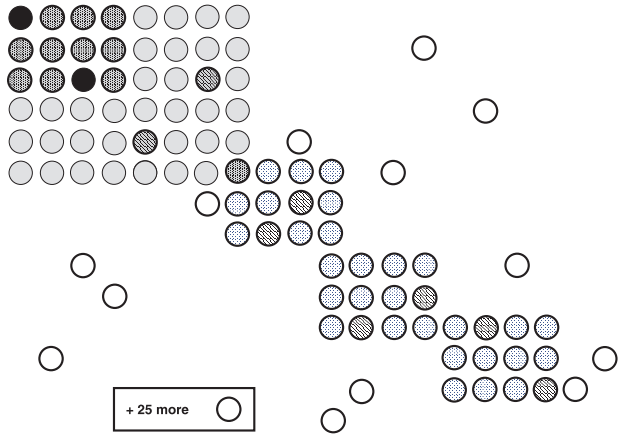



**Figure 2.** Schematic illustration of relationships among 48-member single block, four block, and cherry picked sublibraries analogous to those discussed in the text. Rows and columns indicate individual reagents. Solid circles correspond to products present in all three libraries, whereas open circles represent compounds found only in the cherry-picked library. Gray symbols are included only in the single block set and white stippled circles represent products found only in the four block set. Diagonal hashing identifies products found in the cherry picked sublibrary and one or the other of the block designs, whereas gray stippled circles represent products included in both block designs but not in the cherry picked sublibrary.

*Nearest neighbor profiles*

Direct comparisons of the two extreme designs – *cherry-picked* and *single block* – highlight the complexities of such profiles as well as the inherent asymmetry of such comparisons. When the single block design is taken as reference (Figure 3A), some compounds in the cherry picked design find similar products therein but many do not. Hence that nearest neighbor (*NN*) profile is broad and is displaced to the left, towards lower similarities. In contrast, most compounds in the single block design can find a near neighbor in the cherry picked design, so taking the latter as reference produces a distinctly sharper profile displaced to the right (Figure 3B). This difference is indeed clear from the overall distribution statistics for the two profiles (mean 0.74±0.09 and median 0.72 for Figure 3A *versus* a mean of 0.81±0.09 and a median of 0.80 for Figure 3B), but the relationship is clearly more complex than is indicated by these numbers alone.

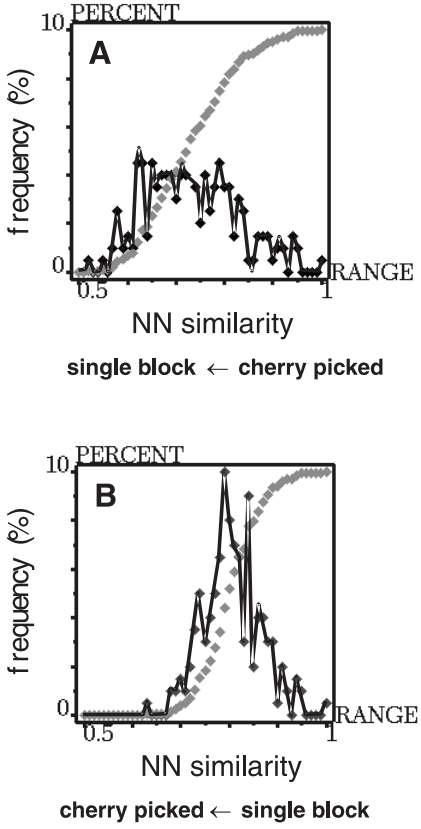



**Figure 3.** Nearest neighbor similarity profiles comparing the cherry picked and single block sublibraries; frequency and cumulative frequency curves are both shown. (**A**) *NN* similarity for products in the cherry picked set, taking the single block sublibrary as reference; (**B**) *NN* similarity for products from the single block design, taking the cherry picked sublibrary as reference.

A somewhat more subtle way to compare two designs is to extract maximally diverse subsets from each and calculate self-similarity profiles for them.[10] Figure 4 shows the results of doing this for the three sublibraries under consideration here, with twenty compounds drawn from

each subset. The cherry-picked sublibrary shows a single large peak at a *NN* similarity of 0.53, with a slight skew towards lower similarities (Figure 4A); such a profile suggests a relatively even coverage. The single block sublibrary exhibits a small peak in this area but has a much larger peak at 0.635, suggesting considerable clumpiness in its spread in structural space. Not too surprisingly, the subset from the four block sublibrary produces a profile falling somewhere in between, with a broad envelope of peaks between 0.525 and 0.57. These differences are even easier to appreciate in the cumulative distributions plotted in Figure 4B. Again, it would be difficult to capture this complexity in any single number.

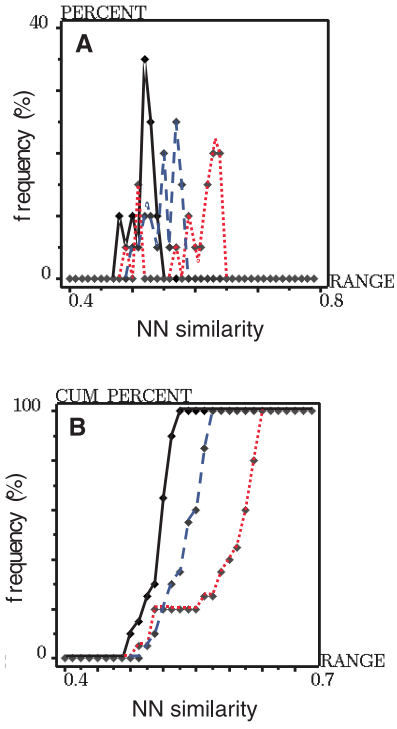



**Figure 4.** *NN* similarity profiles for maximally diverse subsets of 20 compounds drawn from each sublibrary. Solid lines correspond to the subset derived from the cherry picked sublibrary, whereas dashed and dotted lines represent the four block and single block sublibrary, respectively. (**A**) Frequency distributions. (**B**) Cumulative frequency distributions.

### *Fingerprint visualization*

Taking candidate sublibrary comparisons from point (zero dimensional) characterization in terms of a single diversity measure to the one-dimensional *NN* profiles clearly increases the level of useful detail in the information conveyed, but it is still limited. Perhaps most importantly, *NN* similarity profiles only support pairwise comparisons. By moving to *NLM-H* projection into two dimensions (Figure 5), it becomes possible to see how all three sublibraries relate to one another simultaneously.

The *NN* self-similarity profiles in Figure 4 suggested that the single block products were more unevenly distributed than were those from the cherry picked sublibrary. This suggestion is confirmed and elaborated in Figure 5, where it is clear that all three sublibraries "cover" the extremes of the combinatorial space with respect to the distinctive sulfonylthiophene region at the upper left of the projection and the sulfonyl(di)azole area at the bottom. Distribution within the central mass of phenylalkylaminobenzenesulfonamides is similar as well, though the single block library exhibits significantly more clumping; these products make up the bulk of those possible simply because the respective reagent classes dominate those which are commercially available. The single block coverage is more clearly inadequate for the alkylamino alkylsulfonamides, which fall to the right in this plot.

Coverage by the four block design is generally similar to that obtained by cherry-picking products and is clearly better than that seen for the single block sublibrary, particularly in terms of reduced redundancy. Indeed, the only area where the four block library is seriously deficient is in the area enclosed by the ellipse in Figure 5, a deficiency which could be corrected by incorporating a few products from the cherry picked sublibrary, or by increasing the four block diversity by re-running the design program using a larger subsample size *k*.

In one respect, the four block design is actually superior to the cherry picked sublibrary. The latter generates eight products that fall into the singleton "hedge", whereas the former only generates two. Such structurally isolated compounds ("outliers") are generally undesirable candidates for screening, because their activity (or inactivity) is unlikely to be useful in formulating the structure-activity relationships (SAR) required for effective lead follow-up.

## Conclusions

Different design strategies will almost inevitably generate different sublibrary designs when applied to a single combinatorial parent library, with many possible solutions exhibiting similar if not identical aggregate properties. Nearest neighbor fingerprint similarity profiles provide a quick and easy way to characterize the overlap between candidate designs, whereas self-similarity profiles for maximally dissimilar subsets provide a useful way to compare design spreads and coverage for individual sublibraries. Non-linear mapping of fingerprint similarities incorporating an horizon (*NLM-H*) can provide greater

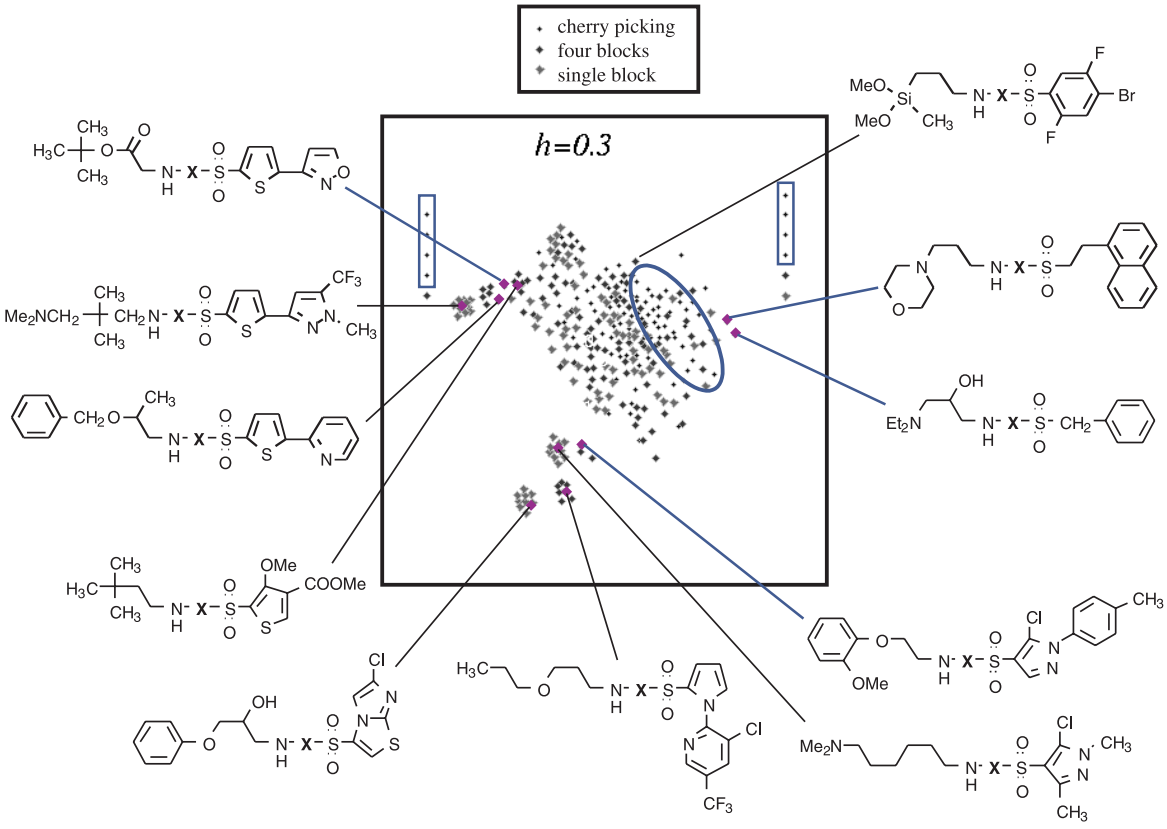



**Figure 5.** Nonlinear mapping projection of pooled sublibraries from fingerprint space into two dimensions based on Tanimoto similarity and a dissimilarity horizon 0.30. The first two principal components were taken as starting coordinates. The **X** in the representative products shown corresponds to the shared carboxamidopiperidine core. The small symbols and large black symbols represent cherry-picked products. The large gray and medium sized symbols represent products from the single and four block sublibraries, respectively.

insight when making more detailed multiway comparisons.

It is generally the case that a cherry-picked design will cover more structural space than will a fully combinatorial (*i.e.*, single block) one.[4] Multiblock designs will necessarily display intermediate coverage, approaching cherry-picked sublibraries as the number of blocks increases. At least for the ureidopiperidine sulfonamide library considered here, application of the analytical tools described here showed that an OptiSim-based four block design using a total of 52 reagents was able to afford qualitatively similar coverage to that obtained by cherry picked products, which would require 400 reagents to synthesize. A similarly obtained single block design, though its realization would require significantly fewer (30) reagents, was clearly inferior to the four-block design in terms of both redundancy and coverage of structural space.

## Acknowledgements

This work was supported in part by SBIR grant 1R43GM58919 from the National Institutes of Health and by Tripos, Inc. Farhad Soltanshai provided helpful discussion and critical technical support.

## References

1. See, for example: Ghose, A.K.; Viswanadhan, V.N. eds.; *Combinatorial Library Design and Evaluation*, Marcel Dekker, Inc.: New York, 2001; Ferguson, A.M.; Patterson, D.E.; Garr, C.D.; Underiner, T.L.; *J. Biomol. Screening* **1996**, *1*, 65; Zheng, W.; Cho, S.J.; Tropsha, A.; *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251; Zheng, W.; Cho, S.J.; Waller, C.L.; Tropsha, A.; *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 783; Sheridan, R.P.; SanFeliciano, S.G.; Kearsley, S.K.; *J. Mol. Graphics Mod.* **2000**, *18*, 320; Martin, E.J.; Hoeffel, T.J.; *J. Mol. Graphics Mod.* **2000**, *18*, 383; Waldman, M.; Li, H.; Hassan, M.; *J. Mol. Graphics Mod.* **2000**, *18*, 412; Mason, J.S.; Beno, B.R.; *J. Mol. Graphics Mod.* **2000**, *18*, 438; Stanton, R.V.; Mount, J.; Miller, J.L.; *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 701.
2. Gillet, V.J.; Willett, P.; Bradshaw, J.; Green, D.V.S.; *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169; Agrafiotis, D.K.; *IBM J. Res. Develop.* **2001**, *45*, 545.

3. Reynolds, C.H.; Tropsha, A.; Pfahler, L.B.; Druker, R.; Chakravorty, S.; Ethiraj, G.; Zheng, W.; *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1470.
4. Gillet, V.J.; Willett, P.; Bradshaw, J.; *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731; Jamois, E.A.; Hassan, M.; Waldman, M.; *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63.
5. Clark, R.D.; Patterson, D.E.; Soltanshahi, F.; Blake, J.F.; Matthew, J.B.; *J. Mol. Graphics Mod.* **2000**, *18*, 404.
6. *UNITY,* version 4.1; Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 USA, 2000.
7. Clark, R.D. In *Combinatorial Library Design and Evaluation;* Ghose, A.K.; Viswanadhan, V.N., eds.; Marcel Dekker, Inc.: New York, 2001, p. 337; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644.
8. Willett, P.; Winterman, V.; *Quant. Structure-Activity Rel.* **1986**, *5*,18.
9. Lajiness MS. In *Structure-Property Correlations in Drug Research*; van de Waterbeemd, H., ed.; Academic Press, Austin,1996, p 179.
10. Cramer, R.D.; Patterson, D.E.; Clark, R.D.; Soltanshahi, F.; Lawless, M.S.; *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010.
11. US Patent applied for.
12. Clark, R.D.; *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181; Clark, R.D.; Langton, W.J.; *J. Chem. Inf. Comput. Sci.* **1998**, *38,* 1079.
13. Patterson, D.E.; Cramer, R.D.; Ferguson, A.M.; Clark, R.D.; Weinberger, L.E.; *J. Med. Chem.* **1996**, *39*, 3049; Brown, R.D.; Martin, Y.C.; *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572.
14. Sammon, J.W.; *IEEE Trans. Comput.* **1969**, *C-18,* 401; Kowalski, B.R.; Bender, C.F.; *J. Am. Chem. Soc.* **1973**, *95*, 686; Hudson, B.; Livingstone, D.J.; Rahr, E.; *J. Comput.-Aided Molec. Design* **1989**, ***3,*** 55; Domine, D.; Devillers, J.; Chastrette, M.; Karcher, W.; *J. Chemometrics* **1993**, *7,* 227.
15. Sello, G.; *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 691.
16. Soltanshahi, F.; Patterson, D.P.; Reiling, S.; *SARNavigator;* Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 USA, 2000.
17. *SYBYL Molecular Diversity Manager,* version 6.6; Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 USA, 2000.
18. *Available Chemicals Directory;* MDL Information Systems, Inc., 146000 Catalina Street, San Leandro CA 94577, 1999.

*Received: April 11, 2002*
*Published on the web: November 8, 2002*