Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods

Dora Schnur

Pharmacopeia, Inc. CN5350, Princeton, New Jersey 08543-5350

Received August 3, 1998

Generation of large libraries of small molecules has required a reexamination of the methods chemists use to select their starting materials for synthesis. When a few reagents can be used to generate vast numbers of compounds, selection of an appropriately diverse set of starting materials is important. Since syntheses involve reactions between specific functional groups, it is reasonable to sort reagents by these groups. Then the diversity of the fragments attached to the given reactive functionality may be examined. This diversity may be defined in terms of the biologically relevant properties of a three-dimensional structure. Cell-based methodology can be used to divide reagents into convenient subsets from which representative diverse reagents can be selected for library synthesis. For ECLiPS libraries, analyses of the reagents used in each individual step have proven to be a useful strategy. Further cell-based analyses of the actual libraries in conjunction with biological activity data have shown clustering of actives in the selected diversity spaces.

INTRODUCTION

As synthesis of combinatorial libraries has become a standard part of the drug discovery process, the development of methods for efficiently designing large collections of drug-like molecules has become a key function for computational chemists. Numerous papers have been published in the past few years,[1−9] and most of the major modeling software vendors have produced modules for library design.[10] Diversity has generally been considered to be important for lead discovery, and much method development has focused on the use of commercial and corporate databases to development diversity/similarity assessment and sampling methods.[11−15]

A key consideration in combinatorial library design is that small numbers of synthons yield large numbers of products; thus, it becomes essential to find a means to optimally choose these small building blocks from the large superset of available reagents to achieve the desired level of diversity. While it is intuitively reasonable and can be demonstrated[16] that the analysis of product diversity is more desirable than of synthon diversity, reagent/synthon selection is generally more feasible in terms of scale. Another important factor enters into the decision to design a library via building blocks rather than final products: synthetic feasibility. The cold hard reality of combinatorial synthesis lies, for the bench chemist, not in the actual library synthesis but in developing practical reaction conditions for the entire set of selected reagents and over the entire set of conversions needed for library synthesis. Often a set of reagents selected by an automated computational technique contains either obviously chemically unfeasible compounds or ones that due to the vagaries of Mother Nature simply fail when tested on solid phase. Substitutions have to be made that may modify but not destroy the diversity design. Since tight timelines are an integral part of pharmaceutical development, an iterative process for reagent selection must be fast and easy. Adherence to a very strict formal experimental design is difficult at best, and a means of finding replacement synthons for the design is essential. While optimal diversity may be an aesthetically desirable goal, it is far more pragmatic to strive for a library design whose diversity is defined by synthetic reality. It should also be based on understandable metrics that can be used later for a SAR (or, perhaps, if synthesis is truly problematic, for a reactivity) hypothesis.

What we have done is to provide chemists with a fast simple property-based tool that groups synthons according to common property descriptors. The chemist chooses and adjusts the properties as appropriate for the library design and can manually select the most desirable starting materials. (Automated subset selection is provided as well if the chemist so chooses.) As a result, the library design method incorporates medicinal chemistry intuition, synthetic feasibility, and the ability to make quick substitutions by selecting another compound from the same group. The emphasis is shifted from covering property space to mapping feasible parts of property space. Holes in the mapped property space are either understood to be synthetically unreachable or consciously omitted.

Since library design by this method employs a kind of "hypothesis space", the intended use of the library is an important design factor. Is this a lead discovery library, a targeted discovery library, or an optimization library? The ultimate library size, amount of diversity/property space covered, sampling density of the space required to find actives, and the diversity metrics used to define the space are all determined by the intended use of the library. While the library should cover a region of diversity space, it is assumed that actives for a given target will be clustered together. Depending on the space and the nature of the target, the activity space may be broad or spiky. Clearly, the selection of appropriate metrics to deal with these issues is a difficult problem and beyond the scope of this discussion.

It is, however, obvious that it is also necessary to analyze the diversity of the products in libraries designed using a synthon approach. Since the ability to compare sets of large

10.1021/ci980138p CCC: $18.00
Published on Web 01/06/1999

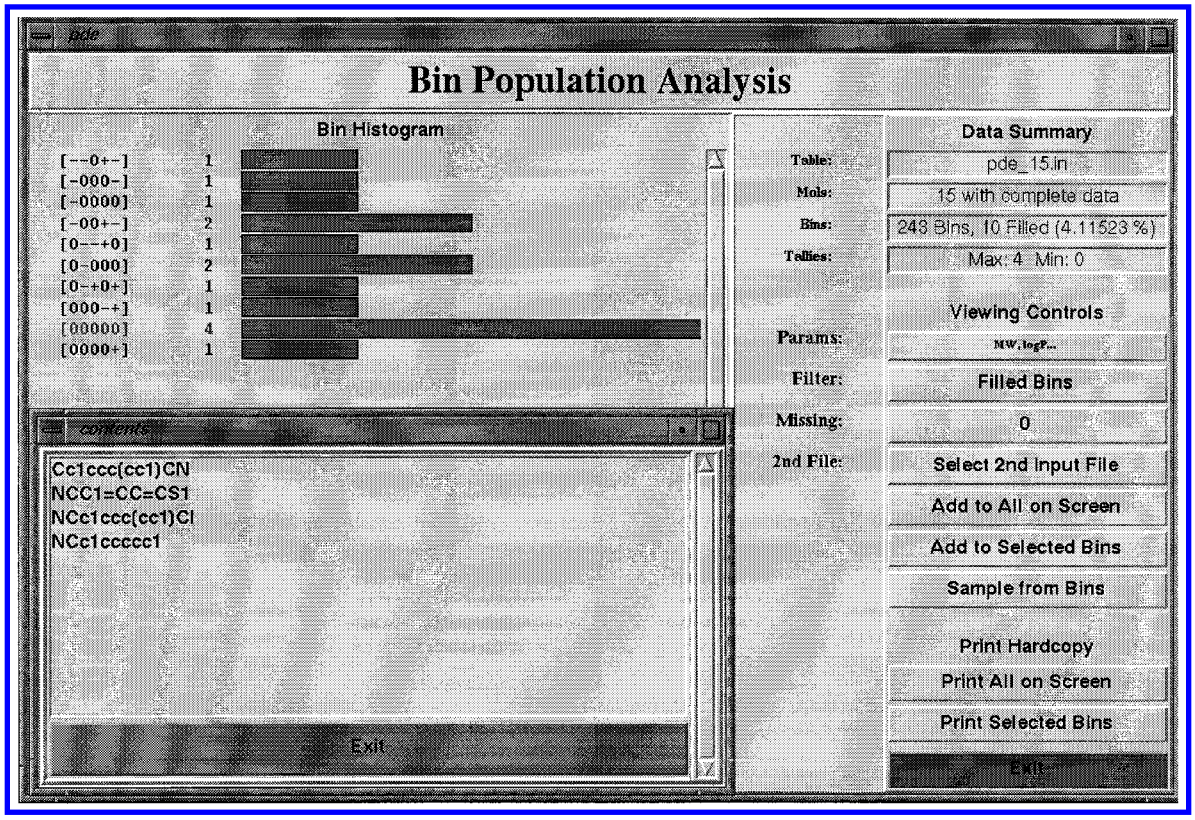

Figure 1. Synthon diversity with bin structures identified as smiles.

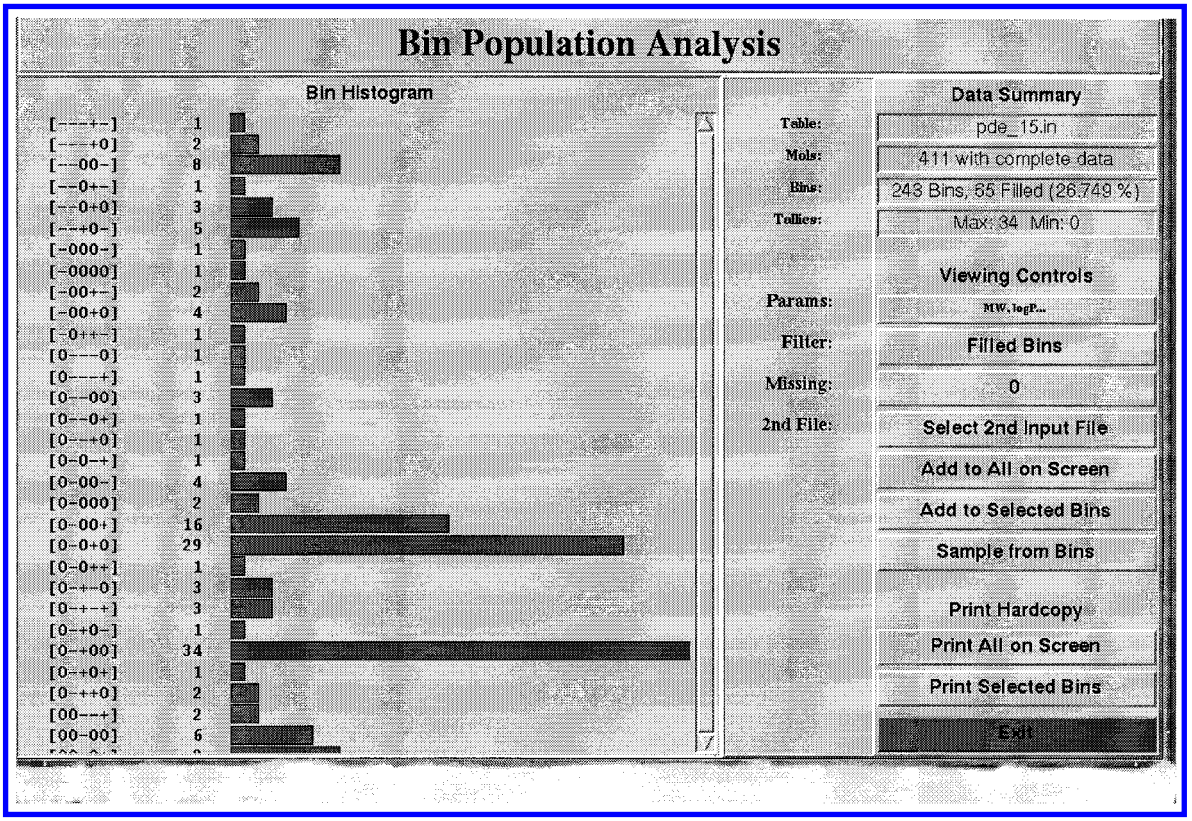

Figure 2. Results of using a complementary data file to fill additional bins.

combinatorial libraries for relative coverage of diversity space was considered essential, a cell-based diversity analysis method was chosen as the most appropriate. Unlike clustering methods, the property space in cell-based methods can be independent of the molecules studied. As a result, molecules can be added or deleted from the set, and different sets of molecules can be compared in the same space. Pearlman's *Diverse Solutions* can easily handle individual analyses of libraries ranging from 20 000 to 150 000 members, can evaluate merged sets of hundreds of thousands of compounds, and was chosen for our analyses. The BCUT metrics[17] used by this program are based on simple physical

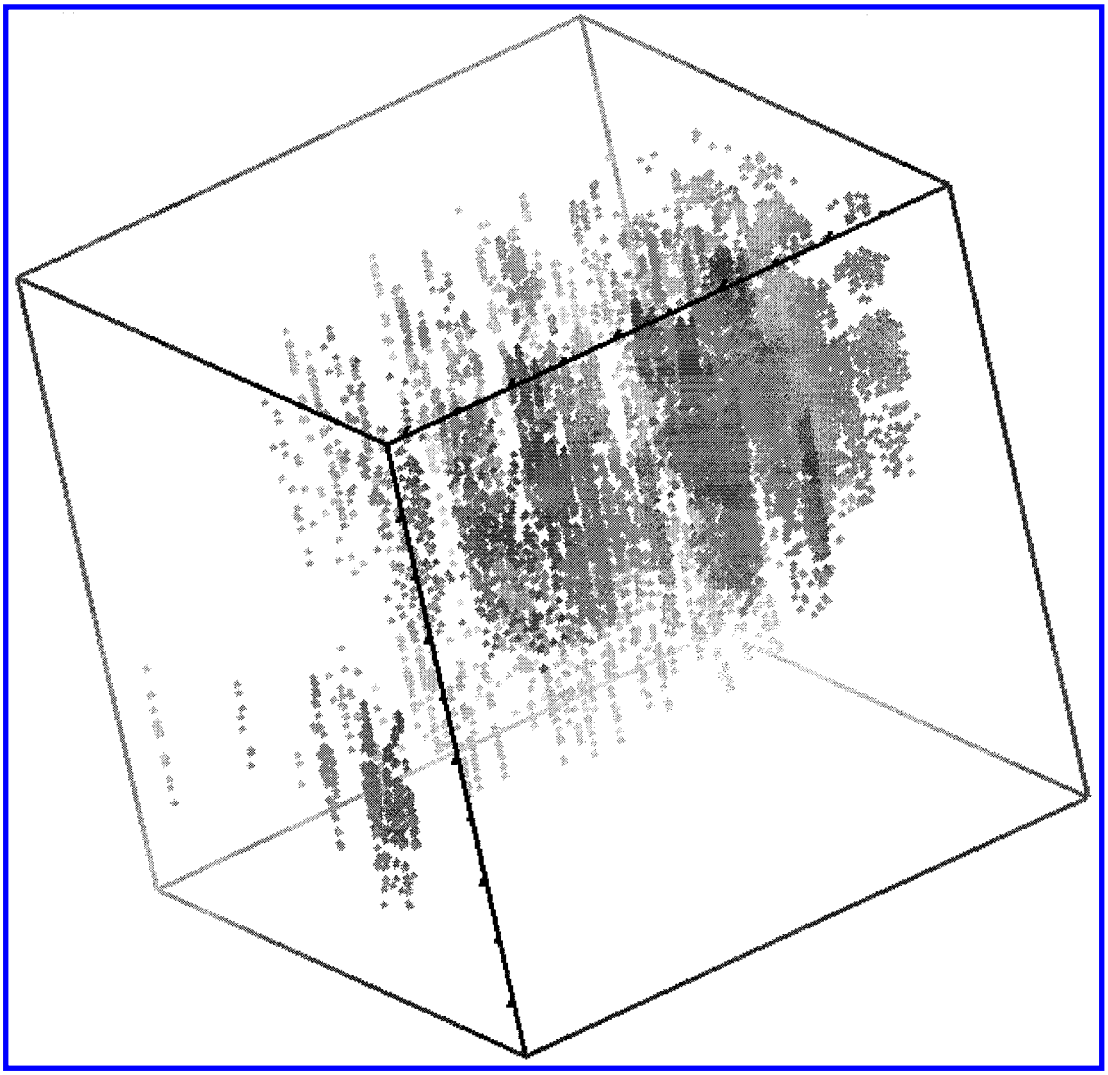

Figure 3. Combinatorial library in BCUT-based diversity space.

properties such as partial charge, hydrogen bonding, and polarizability and should relate to biological activity. We have found this to be the case.

METHODS

Reagent/Synthon-Based Design. Our diversity tool assists in choosing, from among a list of synthons characterized by various molecular properties, a subset in which all feasible combinations of these properties are represented. The basis of this approach is the QSAR assumption that, in the case of potential drug molecules, there is a reasonable number of simple whole molecule properties that may discriminate, alone or in combination, between actives and inactives for a given target.

To make the problem manageable, the range of possible values for each property is subdivided into two or three ranges. For instance, log *P* could be subdivided into low, intermediate, and high ranges, which we will denote by −, 0, and +, respectively. A property such as the presence or absence of a hydrogen bond donor on the molecule can be denoted by just + or −. Any synthon that we might consider using for a given step and which meets all synthetic criteria is thus assigned to a particular "bin". This bin is identified by the property pattern that describes it, i.e., + for lipophilic, − for not having a hydrogen bond donor, 0 for not having an overall charge, or, +−0. These bins correspond to the rows of a full factorial design[18] of either three levels or mixed two and three levels.

Choosing a diverse set of synthons relative to a set of parameters now simply reduces to taking at least one out of each of the filled bins: one from +++, one from ++−, one from +0+, one from +0−, etc. When a bin is empty, the particular combination of parameters may be physically difficult or impossible to realize. Most real world parameters are not completely independent, so finding a small, charged, lipophilic molecule for instance would be difficult. Another more interesting possibility is that the set from which we are choosing our fragments lacks some perfectly reasonable cases and that we should go look for other possible sources—possibly custom synthesis—from which we can fill at least some of these bins.

Our diversity program implements a simple point-and-click interface for choosing a set of parameters and their cutoff values, listing the contents of each resulting bin so the user can select one or more appropriate cases, and automatically filling in bins from other sources of compounds.

Since this tool is designed for bench chemists, a default set of precalculated parameters and ranges for the reagents or other molecules of interest is provided for the user. A good set of parameters is first of all as small as possible,

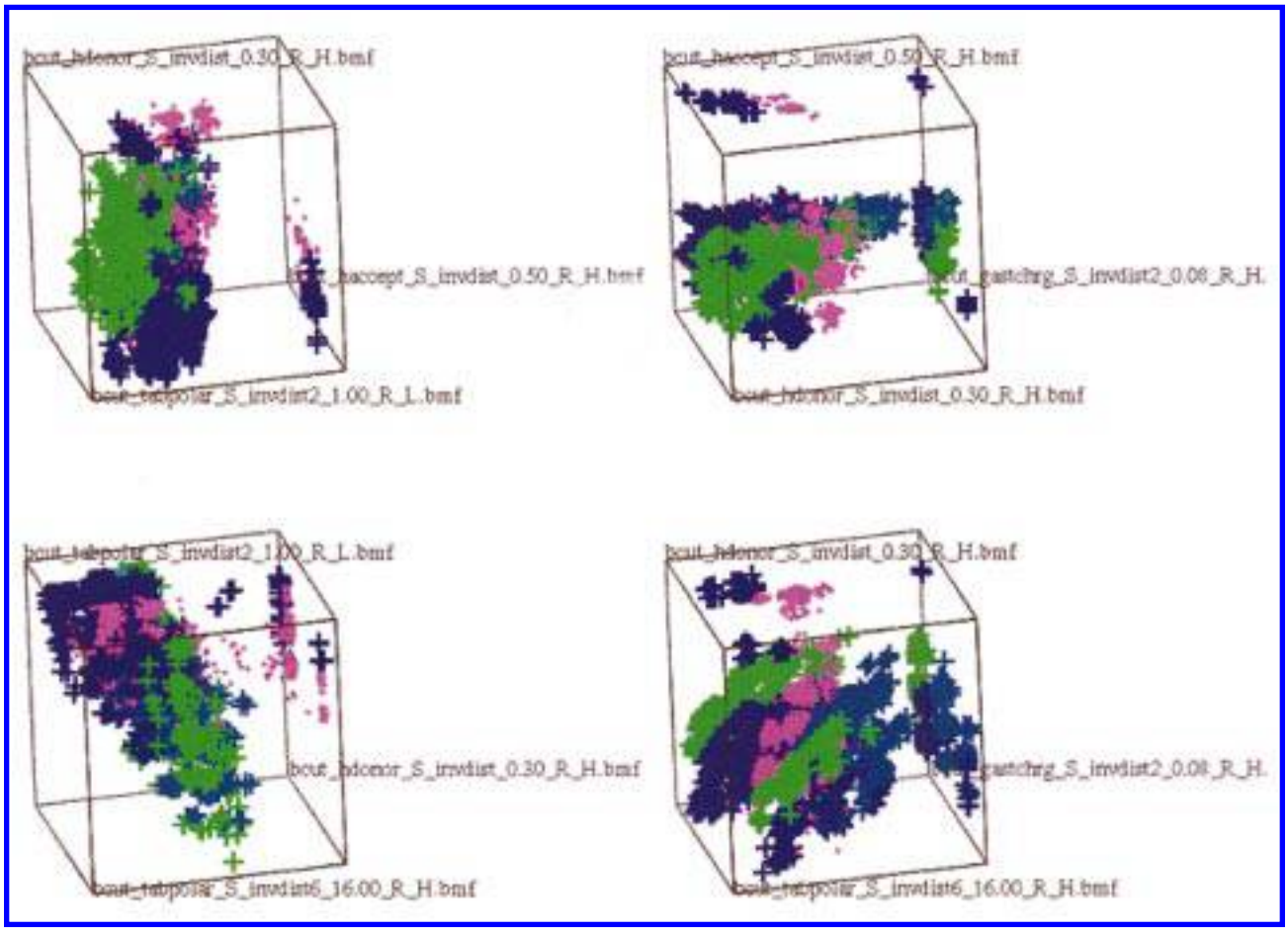

Figure 4. Three libraries compared in four 3D subsets of a 5D BCUT-based diversity space.

since the number of examples we have to choose to represent all possible parameter combinations doubles or triples for each new parameter. Each parameter should represent as much new information as possible. In other words, it should not be correlated with a previously chosen parameter or a combination thereof. So, for instance, we might have available to us molecular weight, molecular volume, and molar refractivity for each molecule. Our experience tells us that size is almost always an important parameter for fragments of bio-active molecules. In any particular example, one of the three parameters mentioned may show a better discrimination than the others, but all three are likely to be correlated to a large extent; thus choosing two of them and varying them independently is both hard to do and difficult to justify. What is justified is the use of combinations of parameters that reproduce a medicinal chemist's intuition (hypotheses) about how molecules should be grouped. A tool that can do this extends the chemist's intuition for sets of molecules too large to be dealt with by visual inspection and, though nonrigorous, will have great appeal to the chemist/user. The key advantage during the library design phase is that the chemist does not exclude medicinal chemistry experience or knowledge of reaction specific feasibility from the process. The interactive approach to setting parameters and selecting synthons from bin groups prevents this.

Program and Output. This software uses a two or three level factorial design as a means of assessing molecular diversity. The input is one or more files containing lists of molecule names with precalculated properties including those from DAYLIGHT,[19] MOLCONNZ,[20] and MOPAC93.[21] Up to nine properties can be selected by the user for analysis. The mean, standard deviation, maximum, and minimum values are automatically computed. These are used to generate thresholds for dividing the property into two or three ranges (+− or +0−). These thresholds may also be altered by the user, if desired. A histogram is then created which visualizes the number of molecules assigned to each property bin. For example, for three parameters in a two level design, the bins would be:

+++
++−
+− −
− − −
−++
−+−
− −+
+−+

If a bin is selected by the user, the molecule names will appear on screen in a new window. Hardcopy for this program includes

*number of parameters

*number of filled bins

*mean, standard deviation, minimum and maximum values, and thresholds for each parameter

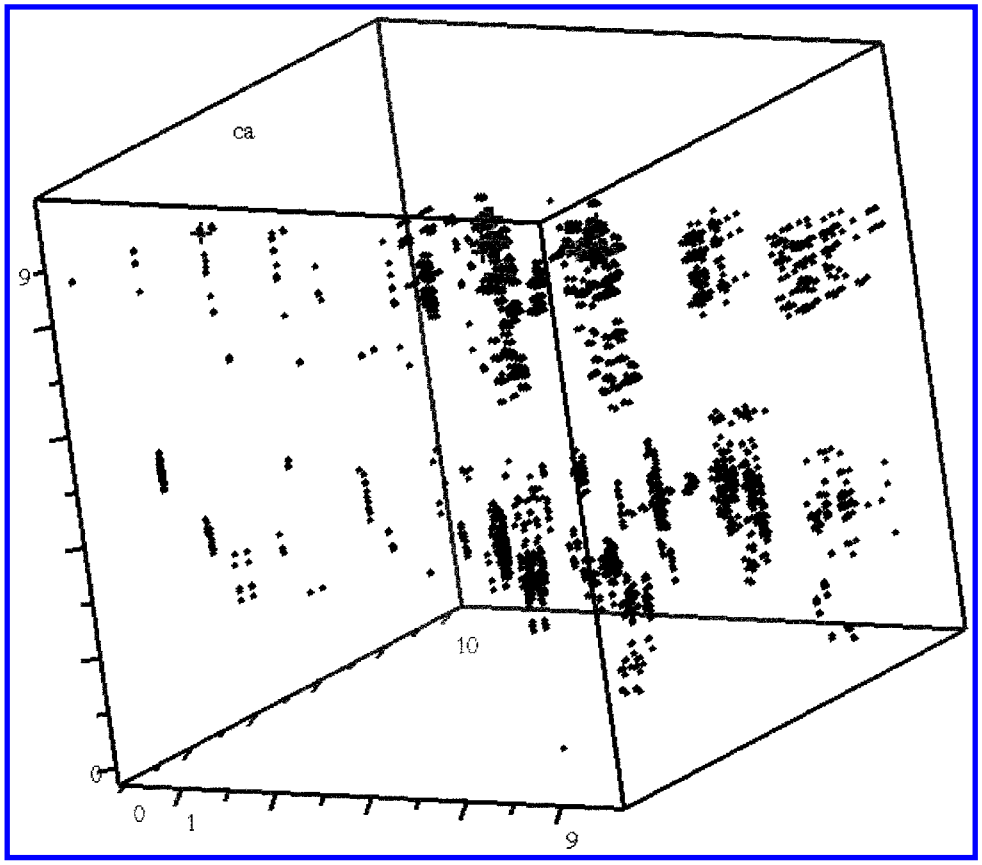

Figure 5. Clusters of carbonic anhydrase actives in diversity space.

*property ranges (values corresponding to +, −, 0) and the associated number of molecules

*histogram of bins with number of molecules per bin

*lists of molecule names associated with each bin

If supplementary files are supplied by the user, a gap or void fill option is available to fill empty or other user-selected bins. These additions are then added to the histogram and to the output bin lists.

Current Applications of This Program. (i) Diversity Analysis of a Reagent List and Reagent Selection. Selection would be performed manually by the chemist using criteria appropriate to the specific design project. Reagent cost could also be used as a parameter in the analysis. Alternatively, a set of more expensive or custom reagents can be used as a secondary file for gap/void fill.

(ii) Diversity Analysis of Model Compounds Derived from Multiple Types of Reagents. Even in a synthon-based approach, it can be of great value to generate a small virtual library of model compounds, particularly in cases where a reaction can use synthons arising from many different types of reagents. The reaction of amines with various carbonyl compounds, alkyl halides, epoxides, etc. is a typical example. In this instance, one amine, such as methylamine, would be used to build model products with each of the different kinds of reactants. Analysis of this product library could be used to decide not only which reagents should be used but also which (and how many of each) reaction type need to be run.

(iii) Diversity Analysis and Comparison of Virtual Libraries. Property-based profiles can be generated, and if the same parameters and thresholds are used, libraries can be directly compared. The property thresholds can also be used as filters to exclude non-drug-like molecules by placing them in the + or − bins.

Crude property/activity analyses can be performed if biological activity is used as a parameter. Selectivity across assays could also be examined.

Diversity Analyses of Products of Large Combinatorial Libraries. For the analyses discussed, various versions of Pearlman's *Diverse Solutions* were used. The methodology and parameters used by this program are discussed elsewhere.[17] Although Pearlman recommends removing 10% of the molecules in libraries analyzed as outliers, all molecules that could be successfully converted from Daylight[19] smiles to 3D structures (SYBYL mol files or MACCS SD) via CONCORD3.2.4[22] were used. While this results in larger diversity or chemistry spaces with more voids, it was felt that all molecules that were actually synthesized and screened should be included in the analysis. Visualization of the resultant chemistry spaces was performed by selecting subsets from the entire library that could be viewed in either SYBYL6.3 (or 6.4)[22] or in Cerius2.[23] In general, subset selection was done by proportional cell-based sampling with a sample size corresponding to roughly 10% of the actual library. The chemistry spaces for entire libraries were visualized using MATLAB.[24] Since Diverse Solutions chemistry spaces range from four to six parameters depending on the size and diversity of the combinatorial library being studied, it is easiest to look at all possible 3D subspaces using the selected subset of molecules. (This assumes the ability to rotate the subspace on a computer screen, an obvious impossibility in a 2D document.) This, in fact, proves interesting when the bioactive structures are added to the subset. (There is no guarantee that any subset picking method

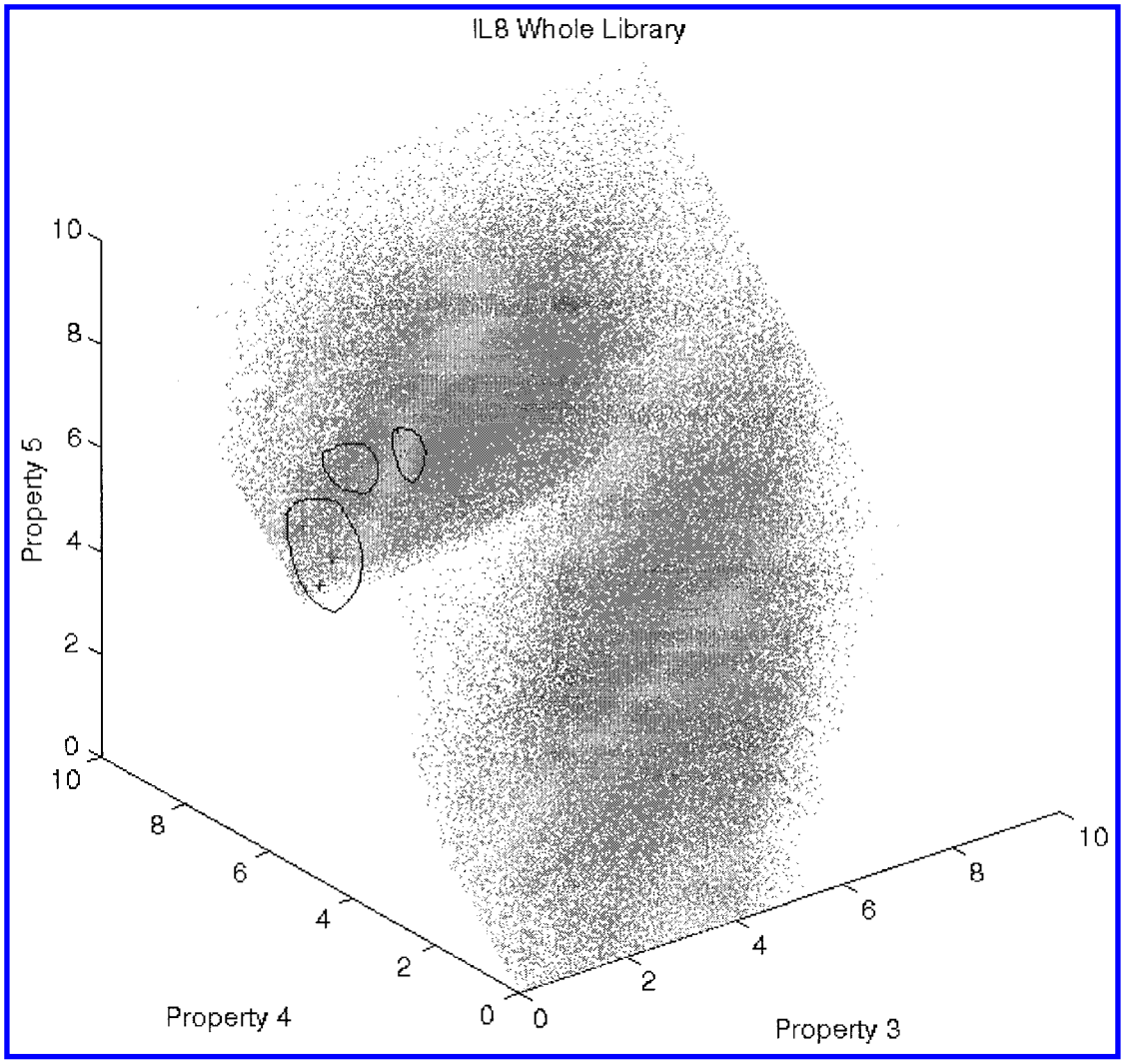

Figure 6. IL8 actives in 150 000 compound library.

will find any, much less all, the actives; so the picked set is manually supplemented.)

DISCUSSION

Figure 1 shows a typical analysis window with a small set of aromatic substituted methylamines binned by their properties. In this case, the chemist chose properties including molecular weight, *c* log *P*, dipole moment, HOMO/LUMO gap, and polarizability for their hypothesis space. (Note that there is no prohibition of using correlated properties if the user so desires.) The secondary window shows the smiles strings for the highlighted bin. Very interestingly, this subset is very acceptable to the medicinal chemist's intuition. The four substituents are 4-methylphenyl, 2-thiophenyl, 4-chlorophenyl, and phenyl. Traditionally, it is assumed that methyl and chloro groups will track similarly for activity (although many exceptions exist) and that the substitution of one sulfur for two aromatic carbons is a standard medicinal chemistry isoster choice.

It was not difficult for the chemist to accept the bin results and replace one or two of the molecules in this bin with more diverse structures in his synthon set. To do so, a secondary file of additional amines was used to fill empty bins with the result shown in Figure 2. The chemist is now able to look at the structures in the additional bins and make selections based on chemical feasibility. Since some of the chosen properties were correlated, not all possible bins were filled, but there are in fact more filled bins than the chemist needed for the total subset. In actuality, not all the filled bins contained amines suitable for the particular reaction and were skipped in the selection process. Had any of the reactions not worked, the chemist could easily go back to the analysis for new selections. In addition, the analysis could be used as a hypothesis for activity-based followup, or the chemist could easily select a new property set for further hypotheses.

Once the various synthons have been selected, it is essential to analyze the resultant combinatorial library in its product diversity space. A well-designed combinatorial library that was designed in synthon space and analyzed using Diverse Solutions is shown in Figure 3. Note that the library is nonspherical and has some outliers. This is expected since it is a multi-core library based on "mix and split" methodology. This is a representative example of a 3D subset of a five-dimensional space, and the axes include BCUT parameters based on Gasteiger charge, hydrogen bond donation, and polarizability.

Comparisons of different libraries in diversity space are of great value, both for designing discovery and optimization libraries. For the former, it is desirable that the libraries fill

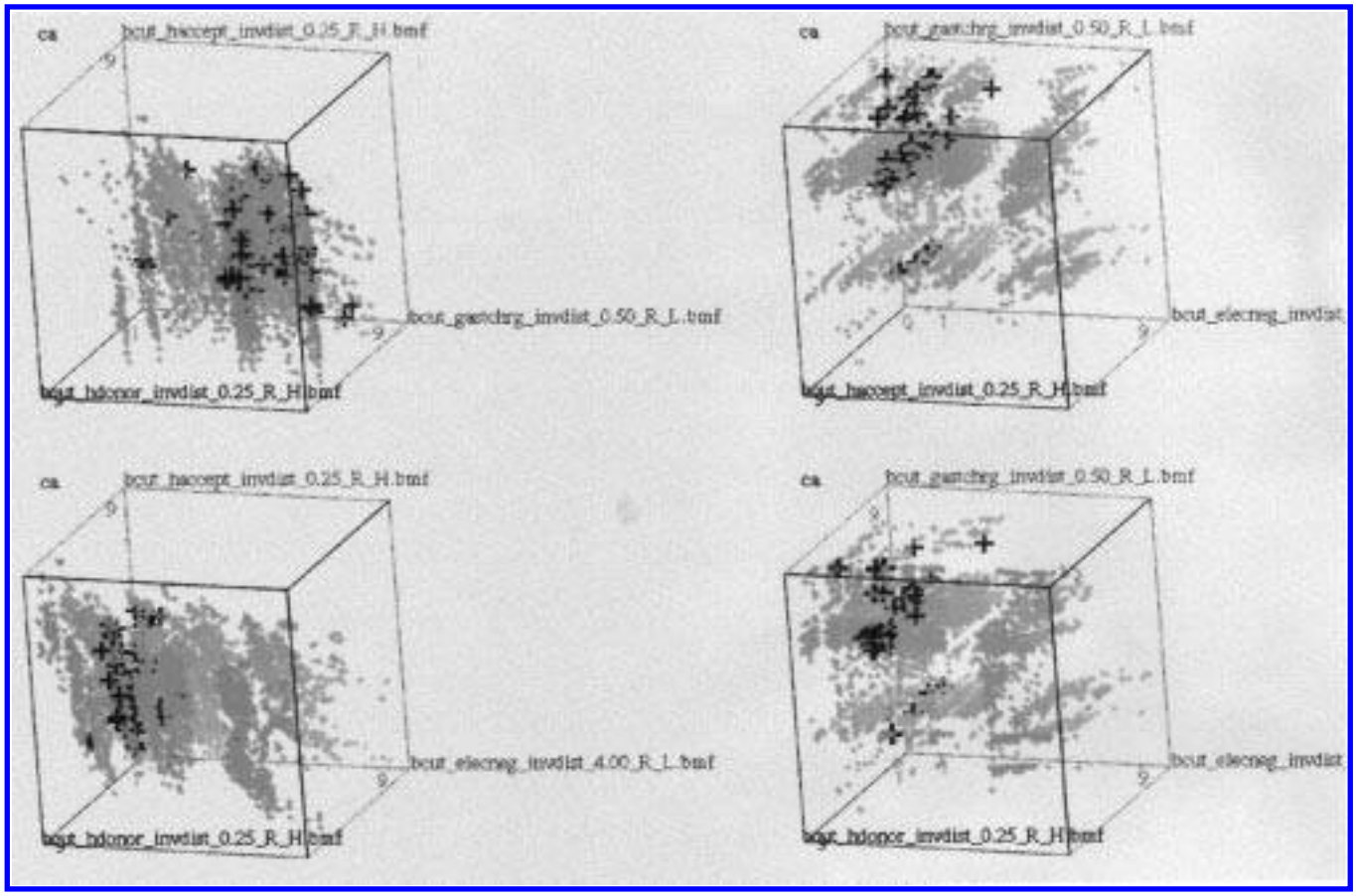

Figure 7. Comparison of library actives for two related targets in activity space.

different cells of the diversity space hypothesis. For the latter, compounds should be concentrated in and near cells where actives are clustered. Figure 4 shows a three library comparison in a 5D space using 3D subsets and BCUTS. Four subspaces are shown.

These BCUT-derived diversity spaces are ultimately useful only if they actually cluster actives. A very early study investigated this with a series of actives from a library designed to inhibit carbonic anhydrase. As shown in Figure 5, there are two major clusters and one minor cluster in this representative 3D subset based on charge, polarizability, and H-bond donation. The result is reasonable-based structure classes studied. Experimentally they fall into three types. Later studies included an analysis of eight actives found for IL8 from a 150 000 compound library. As shown in Figure 6, it was possible to find a 3D subspace that clustered the actives. The worrisome aspect of this analysis is that it is unlikely that any subset picking method on this library would have found these compounds. Assuming that it is reasonable to use diversity spaces as hypothesis spaces for designing active compounds, it is clear from this example that optimal sampling density of a space for a particular target must be considered. It also illustrates the usefulness of large discovery libraries for targets with "spiky" activity profiles. Another test of diversity spaces with regard to activity clustering is the examination of actives for two related targets for which one wants to find selectivity. An example of this is shown in Figure 7. The actives for the two targets (plasmepsin and cathepsin) share common regions of the diversity space but also have nonoverlapping regions.

In the last test of BCUT-derived diversity spaces to be discussed in this paper, five libraries, including three discovery and two focused ones and containing multiple core structures, were compared in a common 6D diversity space with their actives for a particular target. Not only can it be seen (in Figure 8) that the actives do cluster better in some subspaces than in others, but it is also apparent that the more active compounds (Figure 9) cluster more closely. Activity data ranged from ~100 μM to less than 10 nM. To provide further insights into the activity clusters, the BCUTS for the ~90 actives were imported into our factorial design tool, and the parameter thresholds were set so that all actives less than 1 μM were in one bin [−000000]:

[-000000]	***33:##***
[+-0--+-]	**1:#**
[+00-0+0]	**1:#**
[+00-+00]	**3:#####**
[+000-0-]	**1:#**
[+0000-0]	**1:#**
[+00000-]	**1:#**
[+000000]	**30:##**
[+00000+]	**2:###**
[+000+00]	**2:###**
[+000+0+]	**1:#**
[+00+0-0]	**1:#**
[+0+0000]	**3:#####**
[+0+000+]	**1:#**
[++00000]	**4:#######**
[++00+0-]	**1:#**
[++00+00]	**1:#**
[++00+0+]	**1:#**
[++0++0+]	**2:###**

The qualitative message here is clear: while the more active compounds clearly cluster together, there are many less active (but still active) compounds in the region of cell-

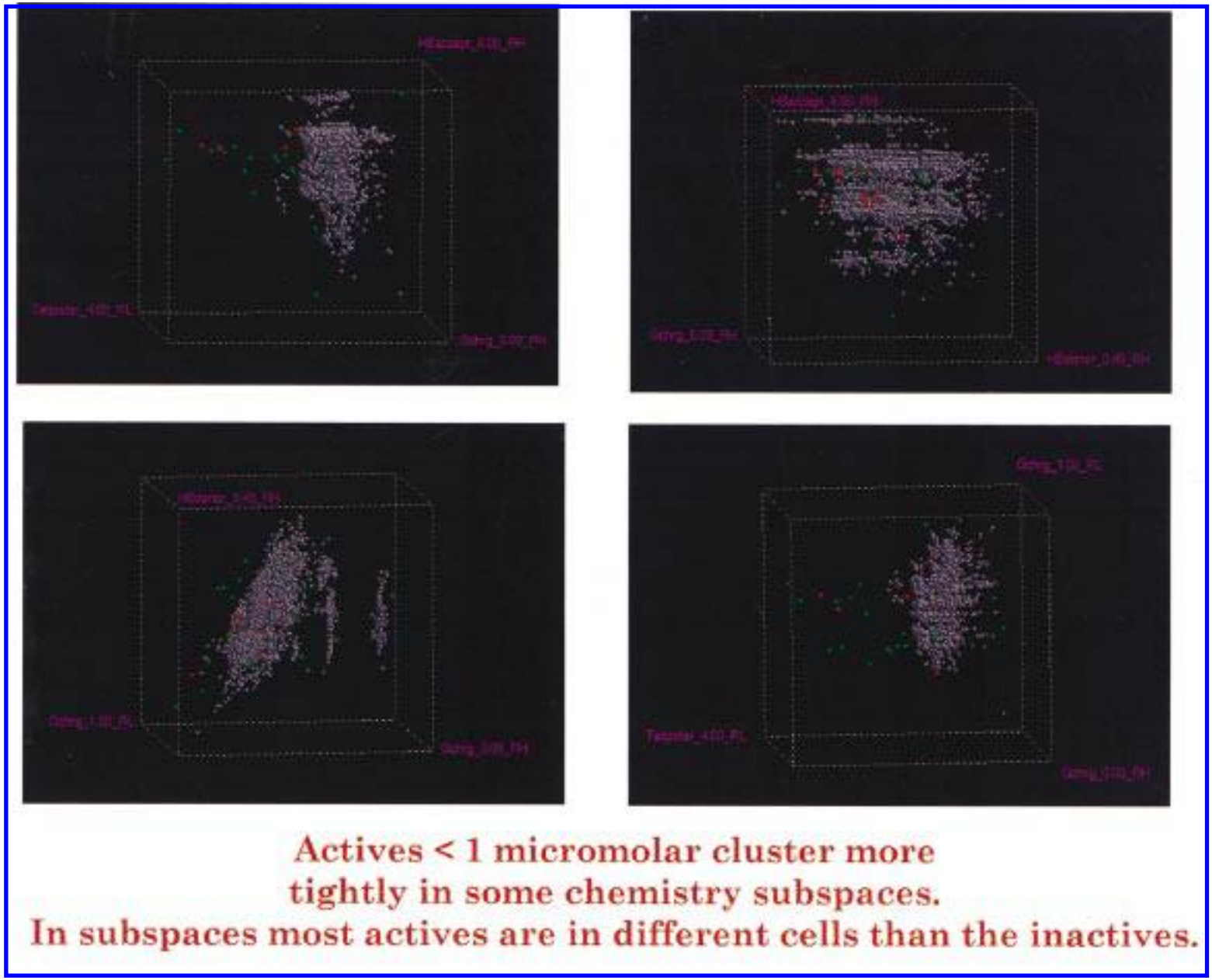

Figure 8. Diversity subspaces for five libraries with actives (red and green).

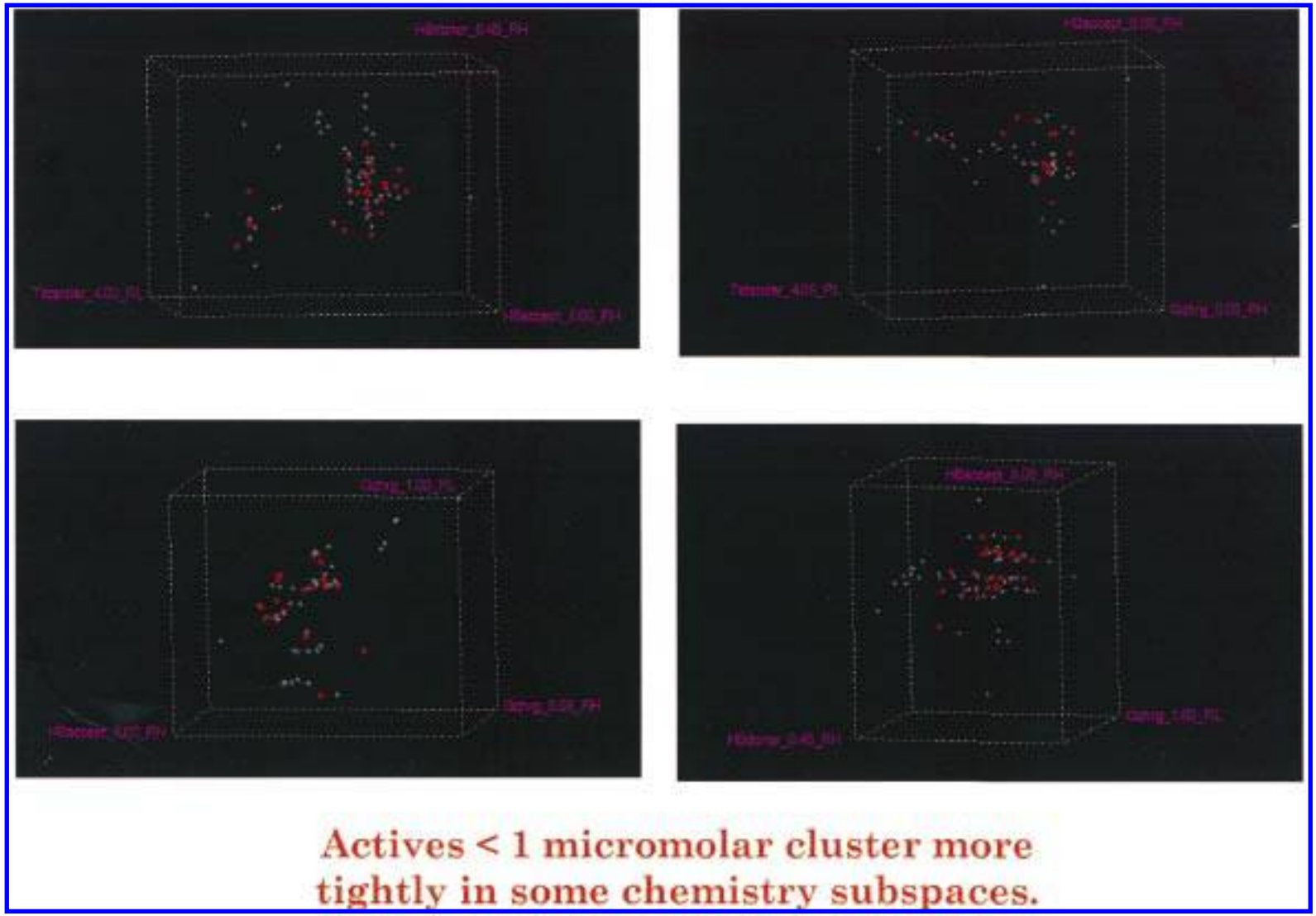

Figure 9. Clustering of multi-core actives with more active compounds (in red) more tightly clustered.

based BCUT diversity space. Clearly, in this instance, the parameters function well for diversity and lead finding/optimization but are not appropriate as QSAR parameters. Actives do cluster, but less active compounds will be found in the same cells.

CONCLUSIONS

We have demonstrated the utility of a simple property-based reagent/synthon selection tool intended for the bench chemist. This tool is designed to bin reagents according to patterns based on the ranges of a set of user-selected properties that form a diversity hypothesis. Since the chemist may select synthons manually from the bin sets, choices may be biased to allow for synthetic feasibility and medicinal chemistry knowledge/intuition. Simple whole molecule property sets have been found that are in concert with the medicinal chemist's intuition and knowledge of molecular similarity.

In addition, we have found it useful to analyze these synthon-designed libraries as full product libraries by cell based using Diverse Solutions, both as individual libraries and for library comparisons. The size of the individual libraries analyzed ranged from several thousand to 150 000 compounds.

Last but not least, active molecules have been found to cluster in various 3D subspaces of higher dimensional diversity spaces. In particular, for a five library analysis, it was found that actives clustered better in some subspaces than others and that more active compounds were more tightly clustered. Since multiple scaffolds were present in the libraries, their proximity in diversity space suggests the possibility of a common pharmacophore for this target. Further work in this area is being pursued, including testing of the subspaces as hypothesis spaces for activity. An important observation about the BCUT parameters used for this analysis is that while they function well as SAR indicators by clustering actives, they are not QSAR descriptors. Less active compounds were found in proximity to more active ones. Further work with other descriptors is underway to determine whether similar behavior is observed.

ACKNOWLEDGMENT

Many individuals contributed to this work, both through helpful discussions and programming. Among them are Bob Pearlman (University of Texas at Austin); Jack Baldwin, George Lauri, Diane Lynch, and Drew Leamon (all of Pharmacopeia); Eric Jamois (MSI); Mark Grieshaber (Monsanto); and Bob Clark (Tripos).

REFERENCES AND NOTES

(1) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(2) Chapman, D. The Measurement of Molecular Diversity: A Three-Dimensional Approach. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 501−512.
(3) Blaney, J. M.; Martin, E. J. Combinatorial Approaches for Library Design and Molecular Diversity. *Curr. Opin. Chem. Biol.* **1997**, *1*, 54−59.
(4) 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.
(5) Good, A C.; Lewis, R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926−3936.
(6) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci*. **1996**, 1214−1223.
(7) 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.
(8) Patterson, D. E.; Ferguson, A. M.; Cramer, R. D.; Garr, C. D.; Underiner, T. L.; Peterson, J. R. Design of a Diverse Screening Library. *High Throughput Screening* **1997**, 243−250.
(9) Agrafiotis, D. M. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 841−851.
(10) For example, Molecular Simulations: *Cerius2 Diversity Manager*; Tripos Associates: *Selector;* Chemical Design: *Chem-Diverse.*
(11) Lewell, X. Q.; Smith, R. Drug-motif Based Diverse Monomer selection: Method and Application in Combinatorial Chemistry. *J. Mol. Graphics Model.* **1997**, *15*, 43−48.
(12) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892−899.
(13) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules form Large Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(14) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(15) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 118−127.
(16) 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.
(17) Pearlman, R. S. Novel Software Tools for Addressing Chemical Diversity. *Network Science*; 1996; http://www.awod.com/netsci/Science/combichem/feature08.html.
(18) Austel V. Experimental Design in Synthesis Planning and Structure Property Correlations. *Methods Principles Med. Chem*. **1995**, *2*, 49−62.
(19) *Daylight Chemical Information Software, ver 4.51*; Daylight Chemical Information, Inc., 18500 Von Karman #450, Irvine, CA.
(20) *MOLCONNZ ver3.10S*; eduSoft, P.O. Box 1811, Ashland VA 23005.
(21) *MOPAC93*; QCPE, Creative Arts Bldg. 181, Indiana University, Bloomington, IN.
(22) *Sybyl6.3*; Tripos Associates, 1699 South Hanley, St. Louis, MO.
(23) *Cerius2 ver 3.0*; Molecular Simulations, Inc., 9685 Scranton Rd., San Diego, CA.
(24) *MATLAB ver4*; Mathworks, Inc., 24 Prime Park Wat, Natlick, MA.

CI980138P