Where Are the GaPs? A Rational Approach to Monomer Acquisition and Selection†

Andrew R. Leach,* Darren V. S. Green, Michael M. Hann, Duncan B. Judd, and Andrew C. Good‡

Medicines Research Centre, Glaxo Wellcome Research & Development, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY U.K.

Received March 17, 2000

Gridding and partitioning (GaP) is a computational method for the classification and selection of monomers for combinatorial libraries. The molecules are described in terms of the pharmacophoric groups they contain and where those pharmacophoric groups can be located in three-dimensional space. The approach involves a detailed conformational analysis of each molecule. This conformational analysis is done within a common coordinate frame, thus enabling the monomers to be compared. The use of a partitioned space is central to this particular application as it facilitates the identification of regions of space which are not well represented by existing compounds. Several ways to extend the use of partitioned pharmacophore spaces are described. Applications of the approach in monomer acquisition and in library design are outlined.

INTRODUCTION

Combinatorial chemistry is now a key technique within the pharmaceutical industry, not only for generating new lead compounds but also in the optimization of those leads to give candidates for drug development. The great interest shown in combinatorial chemistry is due to the way in which large numbers of compounds can be synthesized at low unit cost, and subsequently tested using high-throughput robotic assays. This potential is constrained by three factors. The first limitation is that not all chemical reactions are amenable to robotic synthesis. The second limitation is that the range of monomers for any one reaction scheme is limited. [We use monomer to mean a starting material, reagent, or building block that gives rise to a variation in the structures within the combinatorial library.] The third limitation is that for a particular reaction scheme and a set of available monomers it is not usually possible to physically synthesize all possible products in the virtual library. Rather, it is necessary to select a subset of monomers which when combined together will give rise to the "optimal" library. Some libraries are designed to be as "diverse" as possible; others are designed to be heavily biased toward a particular target or set of targets. The method described in this paper was primarily developed to address the second problem of monomer availability, but it also has some applications in the third area of library design.

Given that the scope of a combinatorial library is so intimately linked to the initial pool of monomers, there is much interest in expanding the range of monomers available for synthesis. One obvious way to achieve this is by purchasing monomers from those offered for sale commercially. However, there is a limit to the molecules that are available in this way. An alternative is via contract or in-house synthesis. It is often the case that noncommercial monomers are more expensive than those available commercially, making it particularly important that an objective method is used to decide which monomers to acquire. The gridding and partitioning (GaP) approach to monomer selection described in this paper is designed to provide such a method. Some of the key features of this approach are that it uses a pharmacophore description of molecules, that it takes conformational flexibility into account, and that it provides a space within which we can compare potential molecules (or sets of molecules) against those molecules which are already available.

The use of three-dimensional (3D) pharmacophores as a descriptor for "whole molecules" has been established by several groups.[1−6] In this approach each molecule is subjected to a conformational analysis. As each conformation is generated, the locations of the pharmacophoric groups within the molecule are tracked. Typical pharmacophoric groups include hydrogen-bond donors and acceptors, the centroids of aromatic rings, acidic and basic groups, and hydrophobic centers. A combination of three pharmacophoric groups constitutes a three-point pharmacophore, which is characterized by the nature of the pharmacophoric groups and the distances between them. The *pharmacophore key* of the molecule is a bit string, each position of which corresponds to a particular combination of pharmacophoric groups and a particular geometry for the triangle which they define. The lengths of the triangle are quantized, typically by dividing each distance into a series of distance bins. These distance bins may correspond to an even distribution (e.g., increments of 1 Å), or they may be nonuniform (to try to achieve equal populations of the bins).[7] At the end of the conformational analysis the "on" bits in the pharmacophore key thus indicate which three-point pharmacophores the molecule can adopt in a low-energy conformation. The use of four-point pharmacophores has also been investigated;[5] the much larger number of possible four-point pharmacophores means that it is common practice just to store those bits that are set "on" rather than the entire binary key.

* To whom correspondence should be addressed. E-mail: arl22958@GlaxoWellcome.co.uk. Telephone: (44) 1438 763383. Fax: (44) 1438 764918.

† Some of this work was presented at the 1999 ACS meeting in New Orleans.

‡ Current address: Bristol-Myers Squibb, 5 Research Parkway, P.O. Box 5100, Wallingford, CT 06492-7660.

10.1021/ci0003855 CCC: $19.00
Published on Web 08/26/2000

A major advantage of the pharmacophore key is that it provides a very natural partitioned space. Partitioning methods provide one way to try to quantify the "diversity" of sets of compounds. In a typical partitioning analysis a number of axes are defined, each of which is divided into a given number of "bins". If there are N axes and each is divided into b_i bins, then the number of cells in the multidimensional space is

$$\text{number of cells} = \prod_{i=1}^{N} b_i$$

Having divided the space into these cells, each molecule can be allocated to a cell according to the values of its properties. It is then a straightforward matter to select a "representative" set of molecules (e.g., by choosing one or more molecules from each cell). Moreover, the empty cells correspond to regions of the space not yet covered; molecules which can occupy these cells might be of particular interest as they enhance the "diversity" of the initial set.

There are various ways in which the partition axes can be chosen. In some cases a set of near-orthogonal descriptors is chosen.[8] In other cases a factor analysis or principal component analysis is used to reduce the dimensionality.[9] A third approach is to use "BCUT" descriptors.[10] One of the drawbacks to partitioning approaches is that the number of cells increases exponentially with the number of axes; this is why it is usual to try and find a reasonably low-dimensional space within which to work. For our purposes, however, we are largely concerned with the one-dimensional pharmacophore keys which represent the three-dimensional pharmacophore distribution.

The pharmacophore key can be used for a variety of other applications. Perhaps the most obvious of these is as a filter to remove compounds which cannot match a particular pharmacophore query, but it can also be used as a similarity metric and as the starting point for a statistical analysis.[11,12]

GENERATING THE GaP DATA FOR MONOMERS

In this work we are particularly concerned with "capping" monomers. These contain a single *attachment group*. The attachment group is the functional group (or atom) which is common to all monomers in a particular set, where the chemical transformation occurs. Common attachment groups are primary alkylamine, carboxylic acid, aldehyde, secondary alkylamine, etc. The pharmacophore key method as applied to whole molecules is attractive but cannot be applied directly to such monomers. The most obvious reason for this is that monomers (especially capping monomers) are typically rather small molecular species which often do not contain much functionality. Thus many monomers do not possess the minimum three pharmacophoric groups required to construct three-point pharmacophores, even if the attachment group is defined as a "special" pharmacophoric point.[5] We also need to be able to cope with the fact that certain types of monomers can also be used in different reaction schemes. For example, primary amines can be used in amide-forming reactions and in reductive aminations.

The GaP scheme uses a 3D Cartesian coordinate space which is partitioned by dividing the space into cubic cells. Each monomer is positioned within this space with the attachment group at the origin and the adjacent non-hydrogen atom oriented along the x-axis. A systematic conformational analysis is performed for the monomer. For each conformation the coordinates of the pharmacophoric groups within the monomer are identified. The cell occupied by each such pharmacophoric group is recorded (a combination of a specific 3D cell occupied by a particular pharmacophoric group will henceforth be referred to as a *pharmacophore cell*). The entire molecule is then allowed to freely rotate about the x-axis, and the cells which each pharmacophoric group "hits" are also recorded (Figure 1). Six different pharmacophoric groups are currently recognized: hydrogen bond donor, hydrogen bond acceptor, a combined hydrogen bond donor/acceptor (e.g., hydroxyl), aromatic ring, acid (i.e., negatively ionizable), base (i.e., positively ionizable). In addition the cells that can be occupied by a heavy (i.e., non-hydrogen) atom are recorded. The output from this analysis is thus a record of the cells into which the monomer can place each of its pharmacophoric groups. Free rotation about the x-axis is used because it is not yet known how the monomer is to be synthetically combined with the rest of the molecule. Indeed, as indicated above, certain attachment groups can undergo different types of chemistry leading to different conformational properties. Thus an amide bond (as would be formed from an amine reacting with an acid) has conformational properties very different from those of a Csp3-Nsp3 bond (as would be formed in a reductive amination). The free rotation effectively provides a superset of the possible cells that the monomer might occupy. In addition to maintaining a record of which cells can be occupied by the various pharmacophoric groups, the algorithm also determines which pairs of cells can be simultaneously occupied by pharmacophoric groups (e.g., "cell 1234 with an H-bond donor and cell 5678 with an aromatic ring"). For six pharmacophoric groups there are thus 21 different pairwise combinations of pharmacophores. A given monomer may of course contain one, two, three, or more pharmacophoric groups; all of the possible pairwise relationships are recorded.

A number of implementation features are worthy of mention. The initial set of potential monomers is typically provided as a file of SMILES strings. Various preprocessing steps are then performed. These include the removal of common protecting groups in order to expose the functionality that would be actually present in the final molecule (for example, *tert*-Bu esters are converted to the corresponding carboxylic acid; fmoc-protected amines are converted to the free amine, etc.). Molecular weight and rotatable bond filters are applied to remove compounds that are considered too large or too flexible according to user preference. These operations are performed using in-house software written using the Daylight toolkit.[13] The monomers that remain after these steps are then converted into a low-energy 3D conformation using CORINA.[14] The conformational analysis is performed using the ChemX suite of programs[15] which has been extended with in-house routines to track the locations of the pharmacophoric groups and to perform the free rotation. A 1 Å cell width is typically employed. The raw output from the conformational analysis indicates which cells are occupied by which pharmacophoric groups and which pairs of cells are occupied by each of the possible pairs of pharmacophoric groups in a given conformation.

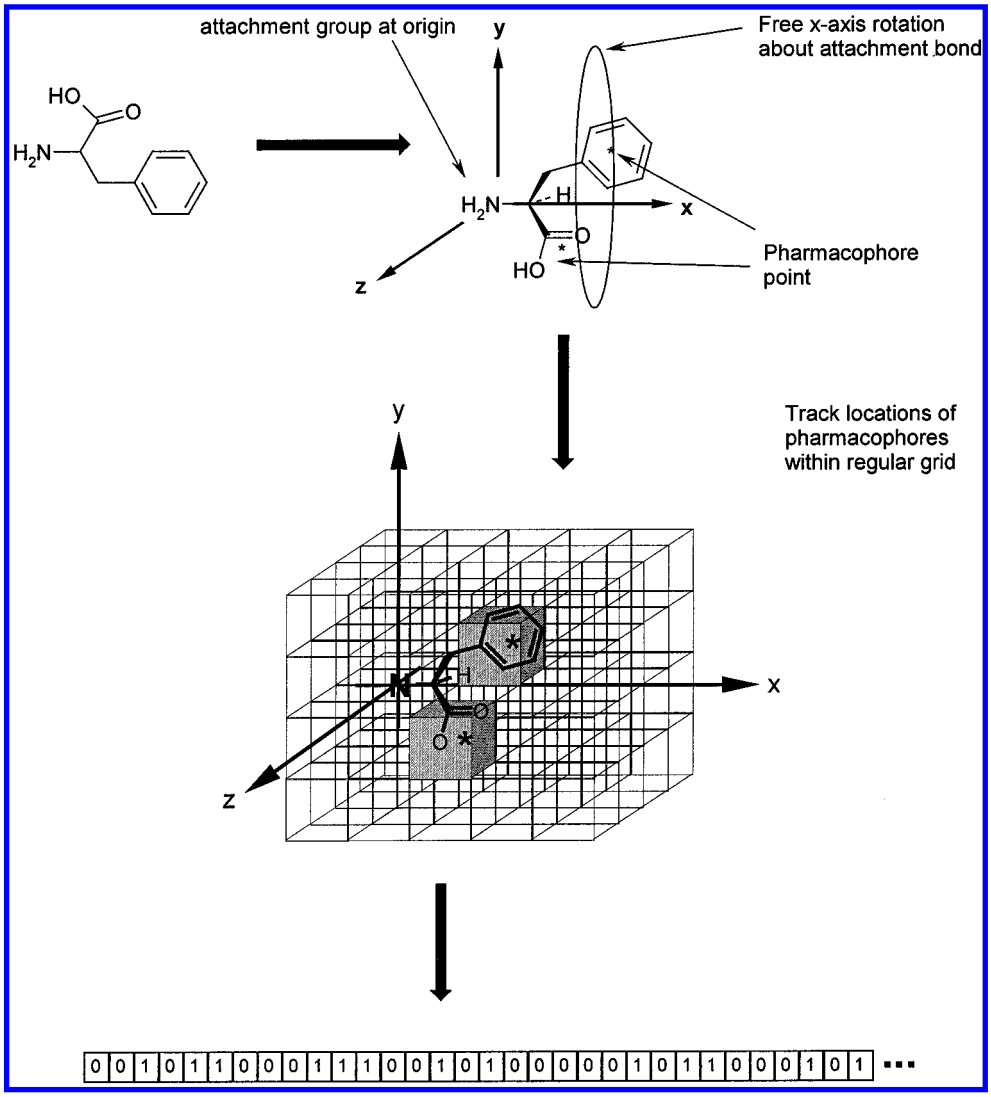

Figure 1. Schematic illustration of the GaP process, illustrated using phenylalanine as an example (primary alkylamine attachment group). This molecule contains two pharmacophoric groups (aromatic ring and acid). The monomer is oriented with the attachment group at the origin and the adjacent bond along the *x*-axis. For each conformation the monomer is permitted free rotation about the *x*-axis, tracking those pharmacophore cells that are "hit".

These data are converted into a compact form for subsequent processing (see below). By way of illustration a graphical representation of the GaP data for phenylalanine is shown in Figure 2, assuming that the primary amine moiety constitutes the attachment group.

While the primary goal for the monomer-GaP project was the analysis of monomers for acquisition purposes, the method has also been extended to cover the situation where we are interested in selecting monomers for a specific library. Under such circumstances the final structures of the molecules in the virtual library are known, so this information can be taken into account. That part of the library which is constant (i.e., the "template") provides the reference point. Moreover, the bond between the attachment group and the template is restricted to torsion angles appropriate to the type of bond. A further extension of this approach is to permit the binding site of a protein to be taken into account during the conformational analysis. This corresponds to having a template docked into the binding site and then performing a conformational analysis on all members of the virtual library that arises from coupling each of the monomers in turn to the protein. Conformations that clash with the protein are discarded. This second variation is closely related to our previously published approach to structure-based library design and monomer selection.[16] A key assumption with both these methods is that the template adopts a common orientation within the binding site.

USE OF GaP DATA IN MONOMER ACQUISITION

There are many ways that the GaP data can be used in monomer acquisition and selection. For the purposes of this section it is assumed that the data generated during the conformational analysis is represented by a bit string (one bit string per monomer). Each position in the bit string corresponds to a particular cell in the 3D gridded space being occupied by a specific pharmacophoric group. Seven pharmacophoric groups (including the heavy atoms), a 1 Å cell width, and a 10 Å box length along each of the three axes would thus give rise to a 7000-element bit string. The actual box size used is chosen to be of a size that ensures that all monomers are contained within it (i.e., about 20 Å long) and so the bit strings generated are somewhat longer.

One of the main advantages of using a partitioned space is that it is very straightforward to compare sets of molecules/

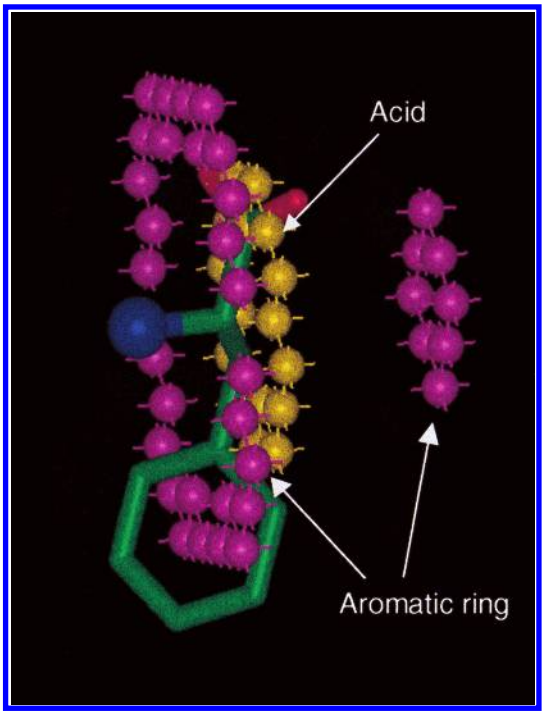

Figure 2. Graphical representation of the cells occupied by phenylalanine. For the acid pharmacophore a single torus of cells is occupied whereas two tori are accessible to the aromatic ring.

monomers and to identify "holes" that one might wish to fill. In the GaP approach we have partitioned the space into just one dimension (i.e., the pharmacophore cell), in contrast to some of the alternative partitioning methods described in the literature which use several properties (or principal components, factors, BCUT values, etc.) as axes, thus creating a multidimensional space. We are particularly interested in monomers which fill regions of the pharmacophore cell space that are not currently occupied by existing monomers. A simple way to achieve this is to first fill the space with the available molecules (e.g., monomers which are already available in-house or which can be purchased at a reasonable price from a commercial supplier). The process of "filling" the space corresponds to performing a logical OR operation on the bit strings for all such available molecules. A simple way to assess each potential new monomer is then to determine the number of pharmacophore cells it can occupy that are not already occupied by any molecule in the available set.

There are various sources of potential monomers. By definition, a potential monomer is any molecule which possesses the appropriate attachment group and is not already on the list of available monomers. The most obvious sources of potential monomers are the various molecular databases of commercially available and/or known compounds. In some cases these may be from specialist monomer suppliers (in which case we would be reasonably confident that we could acquire the molecule should it be selected). In other cases there might be a synthetic scheme associated with the monomer. The third alternative is that the presence of the molecule in the database simply implies that it was synthesized at some stage (and so presumably should be accessible if it is selected). We refer to molecules in the second and third categories as "tangible", as they are presumed to be potentially available. An alternative to sourcing monomers from databases is to construct "virtual" monomers. One way to tackle this problem is as a purely "de novo" approach in which a series of molecular skeletons are systematically generated.[17] A second possibility is to generate virtual monomers using some robust chemical transformation from readily available starting materials. For example, reduction of a nitrile affords a primary amine. This scheme can also be extended to cover virtual molecules generated combinatorially (e.g., secondary amines from reductive aminations).

SCORING FUNCTIONS BASED ON GaP DATA: OCCUPANCY, SECONDARY PARTITIONS, AND NORMALIZATION

A limitation to the use of composite bit strings when dealing with sets of molecules or monomers (particularly when the numbers become significantly large) is that the composite bit string can become saturated. This is particularly the case with the pharmacophore type descriptors where any one molecule may possess a number of pharmacophores (or may be able to occupy a number of pharmacophore cells). One way to deal with this would be to increase the number of partitions by using a smaller cell size. However, this is not warranted by the resolution of the conformational search. Rather, the binary representation is extended to record information about the occupancy of each pharmacophore cell. Thus the number of monomers that can occupy each of the pharmacophore cells is determined. Some desired occupancy level is then defined for each of the pharmacophore cells. For each of the pharmacophore cells that can be accessed by a new candidate monomer, the current occupancy is compared to the desired level. If the current occupancy is lower, then the "occupancy score" for the candidate monomer is incremented by the difference between the desired and actual levels; this gives a greater weight to those monomers which can occupy those cells that are currently poorly filled by comparison with those cells with close to the desired occupancy level.

An additional drawback to the basic scoring algorithm outlined above is that the highest scoring monomers tend to be those with a significant degree of conformational flexibility and with a substantial amount of chemical functionality. This problem has also been observed when using whole molecule pharmacophores.[18] In that case a scaling factor was introduced to penalize molecules that are particularly flexible. We adopt a similar approach for monomers using the following flexibility normalization factor:

$$\left(\max\begin{Bmatrix} N_{\mathrm{rot}} - R_{\mathrm{offset}} \\ 1 \end{Bmatrix}\right)^{-n} \qquad (1)$$

In eq 1, N_{rot} is the number of rotatable bonds in the molecule (a rotatable bond is defined as any nonterminal acyclic, unconjugated single bond which connects two atoms, each of which is in turn bonded to at least one other nonhydrogen atom). R_{offset} is an offset and n is usually an integral power. The functional form of this expression means that, through the use of an appropriate choice of the offset parameter, monomers with up to a given number of rotatable bonds (e.g., two) are not penalized but when the monomer contains more than this threshold value then a penalty is introduced, increasing with N_{rot}. Different values of n are possible; we have often used a value of 3 to reflect the fact that volume is (length).[3]

To score a set containing more than one molecule then the normalization factor is given by the average:

$$\frac{M}{\left(\max\left\{\sum_{M}(N_{\text{rot}} - R_{\text{offset}})^{n} \atop 1\right\}\right)} \tag{2}$$

Here, M is the number of molecules in the set. It is also possible to apply an analogous scaling factor which accounts for both the flexibility and the functionality of a monomer (i.e., its "promiscuity"); in this case the number of cells that are occupied is used as the determining factor, again with an offset and a power term.

INCORPORATING ADDITIONAL DATA

Our primary objective was to develop a conformation-dependent, pharmacophore-based descriptor for monomers that would provide us with a partitioned space analogous to that used for three- and four-point whole-molecule pharmacophores. This is provided by the single- and pairwise-pharmacophore descriptors. However, we also wanted to take other properties into account when selecting compounds. In particular, we routinely calculate a variety of properties, which we collectively refer to as the *profile*.[19] The list of properties included in the profile is currently restricted for reasons of speed to ones that can be determined from the molecular connection table and includes counts of hydrogen bond donor and acceptor atoms, positively and negatively ionizable groups, the number of rotatable bonds, the molecular weight, calculated log *P* and calculated molar refractivity, the "maximal binding energy",[19] the number of potential chiral centers and a simple measure of structural complexity determined by counting the number of bits present in the Daylight fingerprint. This profile is converted to a binary representation by partitioning each property into a number of bins and then allocating bits accordingly. In contrast to the pharmacophore cell bit string, the number of bits set in the profile bit string is constant for each monomer (and equal to the number of properties in the profile).

One of the main problems associated with the use of a partitioned space is that the number of cells grows exponentially with the number of dimensions in the partitioned space. Thus it is often necessary to either limit the number of dimensions or to divide each dimension into a rather small number of partitions. By comparison, the pharmacophore bit string is a single-dimension descriptor, albeit one that is divided into a very large number of partitions. In addition, the effects of conformational flexibility and the presence of multiple pharmacophoric groups means that a typical monomer will set several bits in the pharmacophore bit string (unlike a conventional partitioned space where each molecule is assigned to a single hypercube). To incorporate the extra profile data into the scoring scheme, a hierarchy of scoring values is employed. One hierarchy that we have used to rank potential monomers is as follows:

(1) number of new (i.e., currently unoccupied) single pharmacophore cells filled

(2) number of new profile bins;

(3) pharmacophore cell occupancy score

(4) profile occupancy score

When comparing one molecule to another, the first of the specified values is considered initially. Higher values give rise to higher ranks. However, should the first value be equal for the two molecules, then the values for the second parameter are compared, then the third, and finally the fourth. An alternative to the use of the profile would be to consider the number of new pairwise pharmacophore cells; this provides an overall measure that is more heavily biased toward pharmacophores. As is the case with three-point and four-point pharmacophore for whole molecules, there are usually many more pharmacophore pairs than pharmacophore singles, giving a greater degree of discrimination between potential monomers.

A limitation with the simple occupancy scheme is that the structural "diversity" of a set of molecules is rarely uniform. A typical compound or monomer collection will tend to have an over-representation of certain structural classes. This can be due to many factors, two of the most common being the ease with which analogues can be synthesized and the fact that the compound collections of most large companies reflect the lead series which have been explored in previous medicinal chemistry programs. When the occupancy criteria are applied to these circumstances certain pharmacophore cells may be filled very quickly, but by very similar molecules. While the GaP method is heavily based on 3D pharmacophore information, it is important to take other factors into account. For this reason we have developed the notion of a "secondary partition". This adds an extra dimension to the pharmacophore bit string, similar to the approach taken in the traditional partitioning approaches. Associated with each of the pharmacophore cells is a bit string that represents additional characteristics of the molecule. We refer to this as the *secondary partition*. For example, a structural key or hashed fingerprint may be chosen as the secondary partition, as an approximate measure of structural diversity. An alternative choice would be a bit string that represents some additional properties (e.g., counts of hydrogen bond donors and acceptors, a calculated partition coefficient, etc.) to represent other nonpharmacophore but still relevant features of the molecule. When the reference set of molecules is considered, the logical OR function is now applied not only to the pharmacophore bit strings but also to the secondary partition bit strings, one for each of the pharmacophore bit strings. The score associated with the secondary partition corresponds to the number of new bits set in the secondary partition bit strings, for those pharmacophore cells which are currently underoccupied. This is illustrated for a simple example in Figure 3. The secondary partition therefore provides a measure of the additional structural variation within each pharmacophore cell that a potential monomer may add to the set.

SELECTION METHODS

There are a number of methods that can be used in conjunction with a partitioned space for subset selection. The main thrust of our work was to identify monomers which added most "value" to our existing collection, in terms of the new pharmacophore space that they can access. As noted above, a simple way to identify such monomers is to assign each monomer a score that indicates how many currently unoccupied cells it can access. More sophisticated scoring functions are also possible, as we have discussed. Given that

1	0	0	1	1	1	0	0	1	1	Bit 1
0	1	0	1	1	0	0	1	1	1	Bit 2
0	1	0	1	0	1	1	0	0	0	Bit 3
0	0	0	1	0	0	1	0	1	1	Bit 4
1	0	0	0	1	1	1	0	1	1	Bit 5
2	3	0	6	8	5	3	4	2	7	occupancy

Composite

0				0			1		1
0				0			0		1
1				1			1		0
0				1			1		0
0				0			0		0
1	0	0	0	1	0	0	1	0	1

Occupancy score = (5-2) + (5-2) = 4
Partition score = 1+3=4

1				1			0		1
0				0			1		1
1				1			0		0
0				0			0		0
1				0			0		0
1	0	0	0	1	0	0	1	0	1

Occupancy score = (5-2) + (5-2) = 4
Partition score = 1+0=1

Figure 3. The secondary partition scheme illustrated using a simple situation with 10 bits in the pharmacophore bit string and 5 bits in the secondary partition bit string. The top table indicates the situation for a "reference" set of monomers. The occupancy levels are as indicated in the bottom row, and the values in each of the 10 secondary bit strings are as shown. Suppose the desired occupancy level is 5. There are six pharmacophores that are currently occupied below this level (shown in gray). The second and third tables show pharmacophore and secondary bit string data for two potential monomers together with their occupancy and partition scores. Both monomers have the same occupancy scores, but the first monomer has a higher secondary partition score, primarily because it has a higher number of new secondary bits for the eighth pharmacophore.

it will not necessarily be possible to acquire all such molecules, a selection algorithm is required to identify a subset of monomers to target for actual acquisition. In the sequential selection algorithm, each potential monomer is scored versus the current set. The top-ranked monomer is identified and added to the list, following which the remaining $(N - 1)$ monomers are rescored. The top-ranked monomer is again identified and added to the set, and the remaining $(N - 2)$ monomers are rescored. This process continues until there are no monomers remaining. An alternative is to use a genetic algorithm (GA). Each chromosome in the GA encodes a potential set of molecules. We use a steady-state GA together with crossover and mutation operators.[20]

APPLICATIONS

To illustrate the use of the normalization formula and the secondary partitions, we now discuss some experiments performed using a set of primary alkylamines extracted from the Available Chemicals Database (ACD, version 97.1]).[21] These molecules were chosen to meet certain price and supplier critera. This initial set was pruned to eliminate molecules which possessed undesirable functionality, molecules which had too high a molecular weight (greater than 200), or molecules which had too many rotatable bonds (more than four).[22] A GaP calculation was performed on these monomers. The sequential selection algorithm was then applied using just the total number of pharmacophore cells filled by each monomer as the score, without any normalization. The top 12 monomers from this selection are shown in Figure 4. Thirty-eight molecules were required to fill all the accessible pharmacophore cells. As can be seen, these 12 top-ranking monomers contain some rather flexible molecules which are more akin to "whole molecules" than

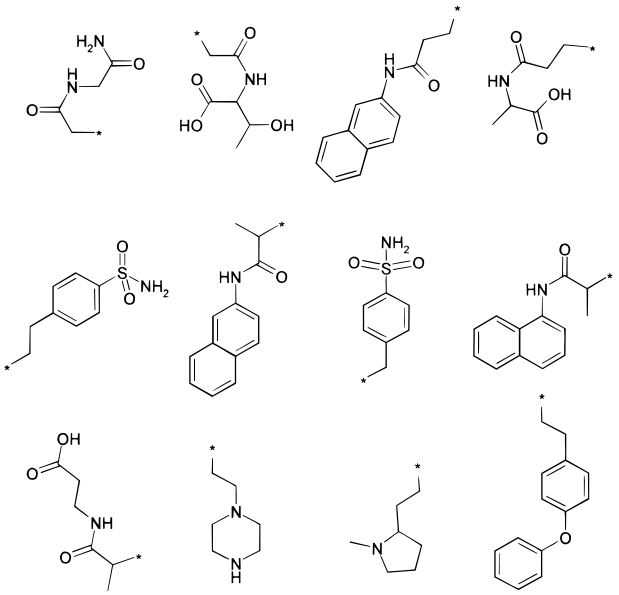

Figure 4. The top 12 monomers obtained from selection of primary amines from the ACD using unnormalized scoring scheme.

capping monomers. A second sequential selection was then performed, using a rotatable bond offset of zero and $n = 3$ (eq 1). In this case 50 molecules were required to fill all of the available cells, and the top 12 molecules from the selection are shown in Figure 5. The top-ranked molecules are now rather more desirable in terms of their ability to act as capping monomers.

We next illustrate how GaP can be used to select a "complement" set of monomers using the occupancy and secondary partition methods. We took as our starting point the 50 monomers selected using the normalized scoring

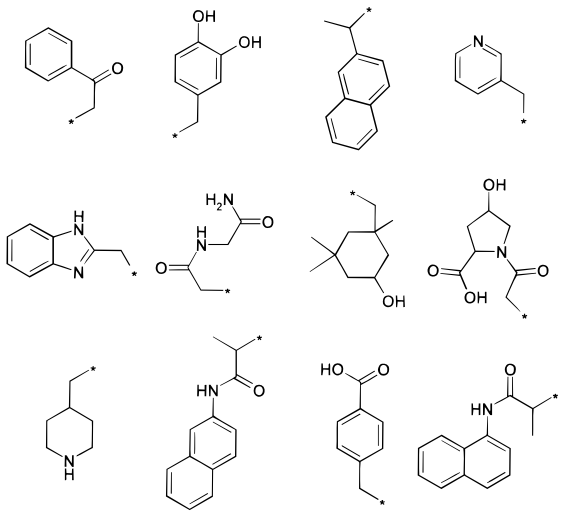

Figure 5. The top 12 monomers obtained from selection of primary amines from the ACD using normalized scoring scheme.

function. This set can access all the available pharmacophore cells with at least one molecule per pharmacophore cell. The sequential scoring scheme was used to select additional monomers, using both the occupancy score and the partition score as described above. For the partition score the hashed fingerprints generated using the Daylight software[23] were used as the secondary descriptor. The objective was to compare the structural diversity of the sets chosen using these two approaches.

The sequential selection algorithm was employed using the same rotatable bond normalization factor. The "structural diversity" of the chosen sets was determined as the average Soergel distance between all $N(N-1)/2$ pairs of molecules.[24] This metric was calculated for the addition of 5, 10, 15, 20, 25, and 30 monomers to the original set of 50. The resulting dissimilarity curves are shown in Figure 6 for the two selection algorithms. As can be seen, selection using the partition-based method gives a small but significant enhancement in terms of the average molecular dissimilarity between the molecules in the set. Of course, in this particular case the total pool of molecules is rather small and in fact there is a degree of overlap between the sets chosen using the two approaches. The use of the secondary partitions is probably more relevant when used to select "whole molecules" in library design or compound acquisition.[25]

A final application of GaP is as a useful way to complement a selection made manually using "chemist's intuition" or some simpler computational procedure. For example, chemists may manually select a set of monomers based upon their own experience or prejudice. GaP can then be used to identify additional monomers which will "fill in" the set and give a more comprehensive pharmacophore coverage of the space. This provides a balance between "medicinal chemistry experience/intuition" and the computational method.

CONCLUSIONS

The Gridding and Partitioning procedure that we have described in this paper has a number of potential uses in the design and synthesis of combinatorial libraries and arrays. Its use of a conformationally flexible, pharmacophore-based descriptor is more directly related to the fundamental processes of molecular recognition than alternative methods based upon the "two-dimensional" molecular graph which only represent the internal framework of a molecule rather than properties relevant to molecular recognition. Moreover, GaP provides a naturally partitioned descriptor that can easily be used to identify "holes" in an existing monomer collection. There are obviously a number of potential limitations associated with the current procedure concerning the quality of the conformational analysis, the pharmacophore definitions, and the grid-space representation. Nevertheless, we believe that when used in conjunction with other computational and noncomputational methods it offers a useful way to make rational decisions concerning monomer synthesis and library design.

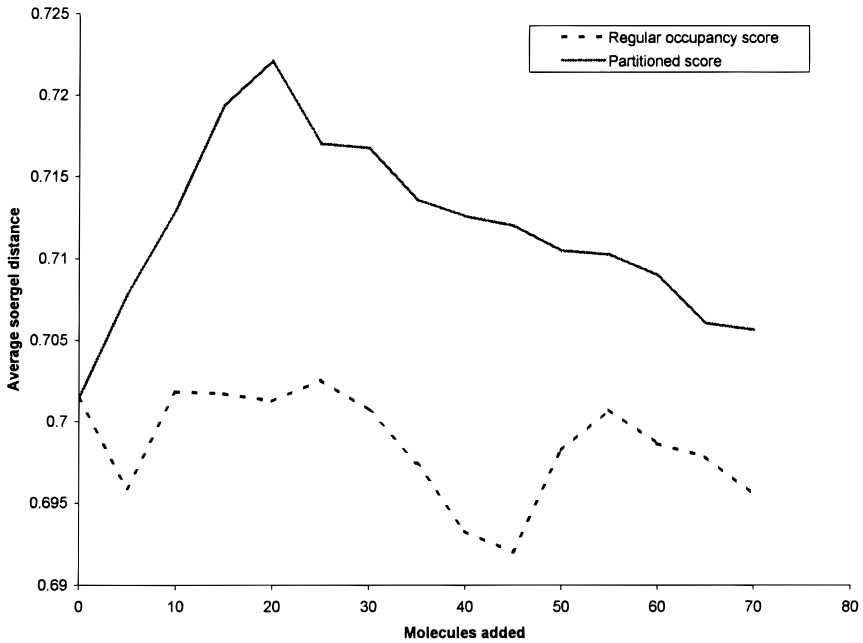

Figure 6. Graph showing the structural "diversity" of monomers selected using the partition and occupancy scoring schemes, measured as the average Soergel distance for all pairs of monomers in the set.

ACKNOWLEDGMENT

We are very grateful to Drs. Ted Baer and Bill Stuart for their support in putting the GaP procedure into practice.

REFERENCES AND NOTES

(1) Good, A. C.; Kuntz, I. D. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9* (4), 373−9.
(2) 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* (3), 572−84.
(3) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB−Strategies for the Design and Comparison of Combinatorial Libraries using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 144−150.
(4) Matter, H.; Poetter, T.. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1211−1225.
(5) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *6* (17), 3251−3264.
(6) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.
(7) Jakes, S. E.; Willett, P.. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* **1986**, *4* (1), 12−20.
(8) Lewis, R. A.; Mason, J. S.; McLay, I. M.. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (3), 599−614.
(9) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (4), 750−763.
(10) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*(3D QSAR in Drug Design: Ligand/Protein Interactions and Molecular Similarity), 339−353.
(11) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (3), 569−574.
(12) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 117−125.
(13) Daylight Chemical Information Systems, Santa Fe, and http://www.daylight.com/dayhtml/doc/prog/prog.toc.html.
(14) Sadowski, J.; Rudolph, C.; Gasteiger, J.. The generation of 3D models of host−guest complexes. *Anal. Chim. Acta* **1992**, *265* (2), 233−41.
(15) ChemX was developed at Chemical Design Ltd and is available from Oxford Molecular, Oxford, U.K.
(16) Leach, A. R. Structure-based selection of building blocks for array synthesis via the World-Wide Web. *J. Mol. Graphics* **1997**, *15* (3), 158−160.
(17) Leach, A. R., Bryce, R. A.; Robinson, A. The synergy between combinatorial chemistry and de novo design. *J. Mol. Graphics Model.*, in press.
(18) 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* (24), 3926−3936.
(19) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J., III. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1161−1172.
(20) Pozzan, A.; Leach, A.; Feriani, A.; Hann, M.. Virtual optimization of chemical libraries using genetic algorithm. Book of Abstracts, 218th ACS National Meeting, New Orleans, Aug 22−26, 1999.
(21) The Available Chemicals Database is from MDL Information Systems Limited, San Leandro, CA.
(22) Note that these filters were applied after the removal of common protecting groups so that it was on the molecule ultimately to be exposed in the final molecule that the filters were performed. It is also worth noting that the total number of molecules in the ACD which contain at least one primary alkylamine is more than 10 000. The number of acceptable capping monomers is thus a small fraction of the total number of potential molecules and illustrates the utility of applying structural filters to sets of potential monomers or compounds.[19]
(23) Daylight theory manual, chapter 6; Daylight Chemical Information Systems, Santa Fe and http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.
(24) The Soergel distance is equal to the complement of the well-known Tanimoto coefficient.
(25) Leach, A. R.; Atkinson, F. L.; Langley, D. unpublished.

CI0003855