Approaches to virtual library design

John H. Van Drie and Michael S. Lajiness

Diversity-based and structure-based approaches to the design of virtual libraries are reviewed. Virtual library design arises in the development of screening libraries, as well as libraries directed at optimizing a particular biological activity. The authors review the thought processes behind different design approaches, and summarize the practical success that some have reported. The major controversies in the field are highlighted, and the assumptions underlying most of these methods are critically analyzed. Reflecting the balance of literature in the field, the majority of this review describes diversity-based approaches, while the rest describes the emerging techniques in 'structure-based combinatorial chemistry'.

With combinatorial chemistry, it is possible to make more molecules than can be tested. Similarly, with the growth of compound collections in pharmaceutical companies, more molecules are available than can be routinely tested. These issues are faced in spite of the enormous advances in high-throughput screening (HTS), and they lead naturally to the topic of 'virtual library design' – the selection of a subset of molecules from a larger library. Virtual library design includes the following activities:

- Selection of a representative or diverse subset of molecules for routine screening from the corporate compound collection.
- Selection of a set of compounds that would best enhance the existing corporate compound collection from external compound collections.
- Selection of the best set of reactions and reagents for the construction of a new screening library, given a series of combinatorial reaction schemes.
- Selection of the appropriate set of reagents from internal or external compound collections to create compounds that optimize that biological activity, given a scaffold for combinatorial chemistry (combichem), with multiple R groups targeted towards a specific biological activity.

Why design libraries?

The first three activities above focus on 'screening libraries', while the last refers to 'directed libraries'. Much of the initial interest in organic combichem was on the development of screening libraries, but increasing attention is now being paid to the combinatorial synthesis of directed libraries. The scope of virtual library design goes beyond combichem, as the first two activities listed above indicate. Although the last two activities listed (i.e. organic combichem) are relatively new, the first two activities have been occurring for over ten years and the experience gained there can enlighten us on the issues confronted with the last two activities.

This review is written primarily for medicinal chemists and individuals engaged in biological screening; secondarily, it is written for all researchers in the pharmaceutical industry. It is not targeted towards our computational brethren; thus, it omits several details on computational issues. The review will focus on work in the original literature on computational aspects of virtual library design. However, space limitations demand selectivity; thus, we do not consider the synthetic methodology of combichem,

John H. Van Drie* and **Michael S. Lajiness**, Computer-Assisted Drug Discovery, Pharmacia & Upjohn, Kalamazoo, MI 49007, USA. *tel: +1 616 833 9302, fax: +1 616 833 9183, e-mail: john.h.vandrie@am.pnu.com

  1359-6446/98/$19.00. PII: S1359-6446(98)01186-6

which has been reviewed elsewhere[1]. Instead, attention is paid to the design of 'organic' libraries, while excluding much of the early work on peptidic or oligomeric combichem and the types of library creation performed in biological systems (such as phage display). This review includes material that has been presented in sufficient detail in the scientific literature and on the Internet; it is surprising to discover the amount of material being presented at conferences that has not appeared in print.

The notion of 'design' under review here is quite different from what is traditionally considered 'rational drug design'. Müller has described this as a paradigm shift towards 'random design and rational screening'[2].

Initial efforts in screening libraries and lead-optimization libraries

The key question in the design of screening libraries is: what is the objective of selecting a sublibrary from a larger library of compounds? Pearlman proposes that 'the obvious objective of that task is to identify a subset which best represents the full range of chemical diversity present in the larger population' (see http://www.netsci.org/Science/Combichem/feature08.html). We tend to prefer objectives that are verifiable experimentally, such as that of Lewis *et al.*[3] – 'any screening strategy should find hits quickly, and furthermore should enable those hit compounds to be turned into a lead series'.

Much of the early work in selecting representative sublibraries descends from the work of Willett and colleagues[4] in the clustering of chemical databases. For many years, Pfizer maintained the structural representatives file (SRF) in an *ad hoc* fashion, which Bawden[5,6] was able to automate via computational methods. At Upjohn, Lajiness[7–9] used chemical similarity and clustering methods to develop the 'Dissim' set, a subset of the corporate compound collection containing representatives of dissimilar groups of compounds. This led to efforts to augment the corporate compound library, based on the criteria that compounds acquired externally should enhance the overall diversity of the library.

The ability to make millions of distinct molecules by combichem first arose in the synthesis of peptides[10] and then peptoids[11]. However, in general, these molecules do not make good pharmaceutical leads and hence such libraries are not particularly useful for pharmaceutical screening. The extension of combichem techniques to organic reactions by Bunin and Ellman[12] and DeWitt *et al.*[13] stimulated an enormous amount of activity in the pharmaceutical industry.

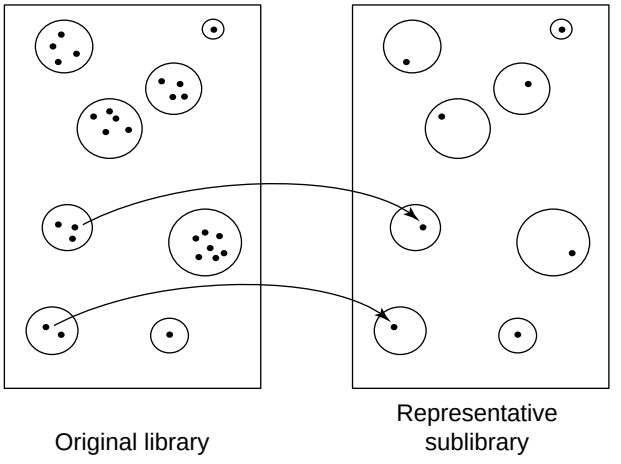

Figure 1. *The overall flow of a generic similarity-based/clustering method in selecting a representative sublibrary from a larger chemical library. Each cluster of similar molecules is circled, and a representative is chosen from each cluster. Notice that outliers (molecules with no neighbors in a cluster) tend to be emphasized by this procedure, as highlighted in Ref. 41.*

Initially, organic combichem was used in the construction of screening libraries, such as the Optiverse™ library from Panlabs and Tripos[14,15]. The design methodology focused on devising a diverse set of organic molecules, based on what the combinatorial synthetic scheme would allow. These methods are reviewed in detail below. Overall, it is possible to classify such approaches into 'similarity-based or diversity-based' methods, and 'grid-based or partitioning' methods.

The design of directed combinatorial libraries for lead-optimization has, until recently, been less well studied. This review examines the emerging approaches to this problem, including the first detailed study of protein-structure-based combichem, a landmark paper from the laboratories of Ellman and Kuntz[16].

Selection methods

Figure 1 depicts the process flow of the similarity-based clustering methods. These begin with a measure of how similar two molecules are. Based on the 'similarity measure', molecules can be laid out in space, clustering those that are similar. A representative subset of the initial library can be created simply by taking one member of each cluster (Figure 1). The precise similarity measures and the

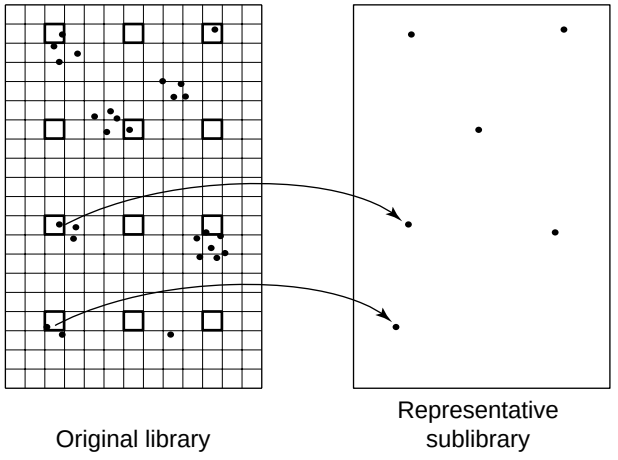

Figure 2. *The overall flow of a generic grid-based method in selecting a representative sublibrary from a larger chemical library. Only one molecule is selected from the central cell of each group of 25 cells.*

clustering methods distinguish the different similarity-based methods.

In grid-based methods, molecular space is divided up into distinct cells, according to a grid (Figure 2). Each molecule in the initial library is assigned a position in the grid, according to its molecular characteristics. Very similar molecules will land in the same cell, or neighboring cells. A sublibrary of the initial library can be constructed by choosing one member of each cell or one member from each group of cells. The types of molecular characteristics used to lay down the grid and the methods used to compute 'coordinates' for each molecule, distinguish the different grid-based partitioning methods. Below are some examples in the literature of each type of method.

Similarity-based selection

The approaches initially pursued by Bawden[5,6] and Lajiness[7] used a non-hierarchical clustering method and a function of the 'Tanimoto coefficient', computed from binary descriptors detecting the presence or absence of specific substructural fragments. The notion of similarity is an attempt to find a mathematical correlation with the chemist's intuitive notion of when two molecules are 'similar' or 'dissimilar'. This is done by assigning a numerical score to the degree of similarity of two molecules, zero for very dissimilar molecules, one for very similar molecules. A variety of methods have been used to measure similarity[4]. The differences between these methods is beyond the scope of this review; however, the fundamental concepts, clustering and similarity measures, are considered in detail below.

Hierarchical clustering is probably the most familiar clustering approach; genomics frequently uses 'dendrograms' to indicate the phylogenetic relationship of a series of related genes. In these cases, the similarity measure may simply be the number of residues in common along the gene sequences.

For the purposes of selecting a sublibrary from a library of molecules, the full hierarchical relationship among all the molecules is not needed, only a set of clusters from which a single member can be selected as a representative is required. Because the number of molecules typically dealt with is in the hundreds of thousands, issues of computational efficiency are important if the selection process is to be performed quickly.

The behavior of such a selection scheme thus depends less on the details of the clustering scheme and more on the similarity measure. Figure 3 illustrates the basic idea of the most common of these, a Tanimoto coefficient computed from two sets of binary substructure descriptors. A standard set of substructures is taken (for example, a phenyl ring, a carboxyl, an isopropyl or a hydroxyl); for

1 **2**

1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

Tanimoto coefficient = A/(A + B + C)
= 3/(3 + 2 + 0) = 0.6

Figure 3. *Construction of a Tanimoto coefficient from a series of substructural descriptors.*

each molecule in the pair, the corresponding bit is set to one if that substructure is present in the molecule, and the bit is set to zero if the substructure is not present. Given these strings of bits for each molecule, it is possible to tally up how similar two molecules are by counting:

- A = the number of bits in common to both molecules **1** and **2** that are set to one
- B = the number of bits in molecule **1** that are set to one, but which are zero in molecule **2**
- C = the number of bits in molecule **2** that are set to one, but which are zero in molecule **1**

The Tanimoto coefficient is simply A/(A + B + C), and measures the proportion of these standard substructures that both molecules have in common.

The example in Figure 3 shows only a few of the many substructural elements present in chemical databases. Extraordinary care was taken decades ago to design a set of substructures for speeding up chemical database searches[17–21], though most workers now simply cite these as 'MACCS-keys'. At Pharmacia & Upjohn, a similar set of substructures is present in the COUSIN chemical database[22].

Alhough the Tanimoto coefficient formula is widely used, it is not without problems – in particular, we have found that it behaves poorly for small molecules. However, other methods have been reported[4] and our standard similarity measures are now composites of the Tanimoto coefficient and other measures based on these substructure bit strings. It is also possible to forego clustering by using a simple heuristic method: pick a random starting molecule; pick a molecule most dissimilar from that; repeat as often as needed[7].

The most serious deficiency of these approaches is that they are strictly '2D', that is, only the connections between molecular groups are taken into account and no reference is made to the molecule's overall 3D structure. But the basis of biological specificity is related to the molecule's 3D structure, hence, it is natural to use this information when computing how similar two molecules are to each other. Cramer *et al.*[23] proposed a novel similarity measure based on 3D bioisosterism. Here, the shape presented by the side-chains of molecules is used to create descriptors that are then used in the similarity computation. Patterson *et al.*[24] compared the performance of these 3D descriptors with 2D descriptors and descriptors based on physicochemical properties, coming to the surprising conclusion that the rank ordering of the quality of descriptors is:

3D with H-bonding = 2D (side-chain only) = 3D steric > 2D (whole molecule) > log P = random numbers.

This issue of '2D vs 3D' remains one of the most controversial issues in this field today (see discussion below).

Instead of clustering, some workers select sublibraries based on a similarity measure using a genetic algorithm (GA). These algorithms mimic Darwinian evolution to optimize the functions of many variables. The key question for the application of any GA to a chemical problem is 'what is the fitness function?', because the purpose of a GA is to find the precise configurations that give the best values for a given fitness function. GAs are especially useful when the number of possibilities to be explored is very large, as is frequently the case in the design of combinatorial libraries.

Brown and Martin[25] adopted an ingenious approach in applying genetic algorithms to selecting a sublibrary from a larger library of possible molecules. In this work, the entities that are evolving are entire sublibraries, and the fitness function is the measure of diversity of that sublibrary. This is assessed along the lines of the 2D and 3D measures described above and, in addition, use is made of 3D pharmacophore-like measures that are described in detail below. In their study, they considered organic combinatorial libraries, one of which had two sites of diversity: R_1 from carboxylic acids, chloroformates and amines; R_2 from carboxylic acids, aldehydes and carbamoyl chlorides. With 360 possibilities for R_1 and 259 for R_2, the total space of possible compounds is 93,240, with a target to select a sublibrary of only 10,000 compounds. After 300,000 generations and 5.5 h of computer time on a 250 MHz Silicon Graphics Indigo2, their GA found the optimally diverse sublibrary. They conclude that this method 'provides a framework for the design of libraries to meet any calculable optimization function'.

The ability to visualize the diversity of a library is useful, as the methods of Agrafiotis (see http://hackberry.chem.niu.edu/ECCC3/papers/paper49/eccc3.html) and Hassan *et al.*[26] demonstrate. Both methods can use any combination of 2D and/or 3D descriptors of the types already described. Agrafiotis uses a Sammon plot to display a molecular library, in a manner that is analogous to a galaxy, where the nearby stars represent similar molecules and stars far apart represent dissimilar molecules. A comparatively dispersed galaxy is apparent for libraries that are inherently diverse, while a compressed constellation is

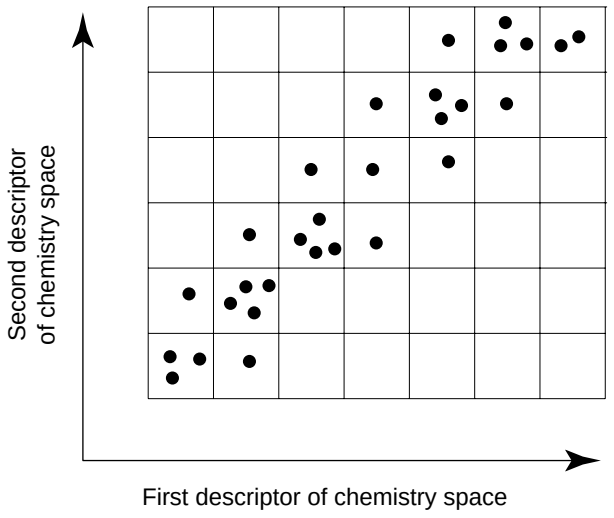

Figure 4. *In grid-based methods, when two descriptors are correlated, artifactual diversity voids tend to appear. Because the descriptor on the y-axis tracks that on the x-axis, it is unlikely to find any molecule which can occupy the upper-left or lower-right quadrant.*

seen for those where the libraries are not very diverse. This type of approach also allows a visual depiction of the degree to which two libraries overlap.

Sadowski *et al.*[27] have described the use of neural networks, Kohonen maps and spatial autocorrelation functions in assessing the similarity and diversity of libraries. In their work, 3D descriptors were computed from the full conformational exploration of each molecule in 49 h on a Sun Sparc 10/512 workstation. Like a Sammon plot, a Kohonen map tends to preserve the property that neighboring molecules are similar and those far apart are dissimilar. This allows the straightforward computation of the overlap between two libraries to decide if the libraries are sufficiently dissimilar to warrant their synthesis.

Grid-based selection (or partitioning)

Two grid-based methods that have been reported use BCUT values, developed by Pearlman *et al.* (see http://www.netsci.org/Science/Combichem/feature08.html), and diverse property-derived (DPD) codes, described by Lewis *et al.*[3] The aim of any grid-based method is to divide up chemical space by laying down a grid, assigning each molecule to a cell in this grid. With BCUTs and DPDs a number of descriptors related to 2D, 3D and physicochemical properties, including charge, polarizability, and H-bonding characteristics, are converted, for example, into six dimensions. A six-dimensional grid is laid out over the space of all molecules, with multiple molecules per cell; to construct a sublibrary, one molecule is chosen from each cell.

There are three advantages of these grid-based approaches over those based on similarity:

- It is a straightforward task to identify 'diversity voids', regions of chemical space that a given library does not cover. Furthermore, it is much simpler, given such a void, of working backwards to what types of molecules would cover this void.
- It is intuitively apparent whether a given library is diverse.
- It is straightforward to assess how similar two libraries are.

One disadvantage of such an approach is that, given a molecule in one grid cell and a void in a neighboring cell, it is not readily apparent what types of structural modifications must be made to the first molecule to create one which would land in the empty cell. Another caveat with grid-based approaches, as Lewis *et al.*[3] have said, is that the final descriptors of the grid must be relatively uncorrelated. Figure 4 shows the effect of a 2D grid using two highly correlated descriptors applied to a very large chemical library. The diversity voids in the corners are actually regions of the grid that are highly unlikely ever to be attained by any molecule.

Lewis *et al.*[3] measured the effectiveness of generating libraries based on their DPD partitioning. In screening for compounds that reduce low-density lipoprotein, one hit was found in the representative sublibrary of their compound collection. Other compounds in the same grid cell were then also found to show activity. DPD-based selection did not yield further activity improvements; however, they developed a pharmacophore query from this information to search a 3D database, yielding two new lead series. They provide one caveat: 'there is a general weakness inherent in all classification and partitioning procedures regarding objects that lie on the periphery… objects that lie close to the dividing boundaries may be misclassified… to avoid missing compounds that just fall outside a partition, the surrounding partitions should also be tested… for a six-dimensional system, with perhaps 50 compounds per partition, this could provide a further potential 36,000 compounds to be screened, which defeats the object of the exercise'.

Mason and Pickett[28] described another partitioning method based on 3D descriptors, the pharmacophore-derived query (PDQ). They considered all triad pharmacophore combinations based on six features (H-bond donors and acceptors, acids, bases, aromatic groups and hydrophobic groups), with the three distance constraints partitioned into the distance ranges 2–4.5 Å, 4.7–7 Å, 7–10 Å, 10–14 Å, 14–19 Å and 19–24 Å. This leads to a total of 5,916 distinct triads that are used to encode a bit string for each molecule (the bit corresponding to each triad is set to one if that triad hits on the molecule, and is set to zero if it does not). These bit strings form the basis for the partitioning of a chemical library. ChemDiverse, as described by Davies and Briant (see http://www.netsci.org/Science/Combichem/feature05.html), appears to be a different implementation of the same conceptual approach.

Another method along these lines is an heuristic algorithm for reagent picking (HARPick), developed recently by Good and Lewis[29]. In this approach, simulated annealing (another optimization algorithm) was employed, rather than a genetic algorithm. Again, the key chemistry aspects are how to describe the space of possible molecules and what function is being optimized. They create combinatorial libraries that maximize the diversity of the library as determined by pharmacophore-type measures.

Structure-based selection

The discussion up to this point has focused on selection techniques in situations where no reference is made to a specific biological target, that is, screening libraries. We now focus on structure-based combichem, where sublibraries of molecules are constructed, biased by structural requirements for activity against a particular receptor.

Wipf *et al.*[30] describe the combinatorial synthesis of a series of serine/threonine-phosphatase inhibitors. They begin by observing that a series of natural products known to inhibit these phosphatases (such as okadaic acid and cantharidin) share a common pharmacophore (Figure 5). This led to the combinatorial design shown in Figure 6, with four sites of diversity. From this combichem library 18 molecules have been tested for biological activity against the protein phosphatases PP1 and PP2A, and the authors reported that 'inhibition has been established in preliminary studies for several members of our library'. Two members of the library were active, based on cellular measurements, with an IC_{50} <100 μM.

Van Drie and Nugent[31] describe theoretical approaches to the problem of structure-based combichem in the absence of the X-ray structure of the receptor. They use DANTE (Refs 32–34) to infer 3D pharmacophores automatically from structure–activity data, which includes information about the shape of the binding cavity (Figure 7),

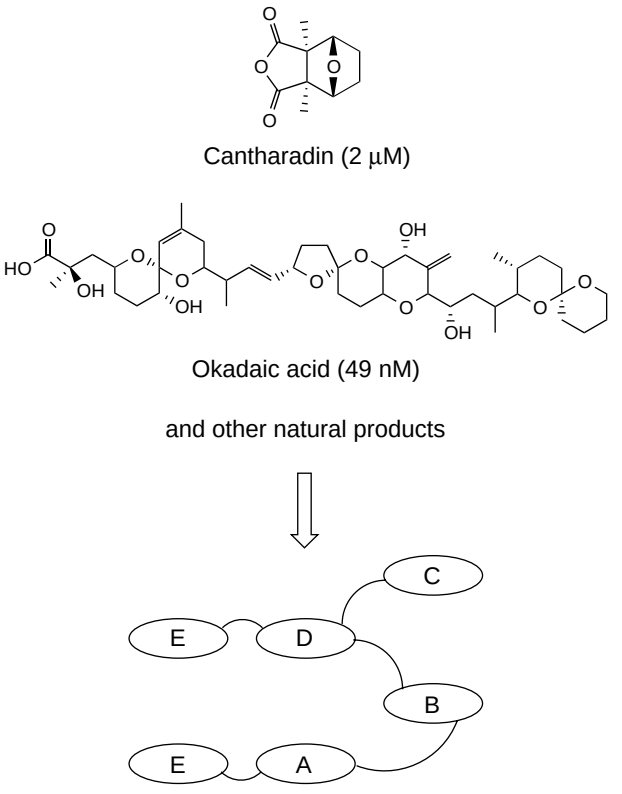

Figure 5. *Pharmacophore common to a series of natural product serine/threonine-phosphatase inhibitors[30]. The domains are: A, H-bond acceptor; B, hydrophobic backbone; C, carboxylic acid, phosphate; D, H-bond acceptor; and E, extended hydrophobic side-chain.*

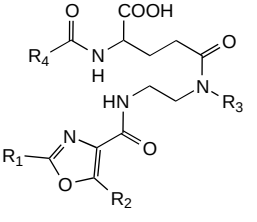

Figure 6. *Combinatorial design based on the serine/threonine-phosphatase pharmacophore in Figure 5 (Ref. 30).*

and search 3D databases for reagents that are complementary to the shape-enhanced pharmacophore. This work introduces the idea of *terra incognita*: regions of the putative receptor site that the SAR has not explored. Searching 3D databases can also be used to look for reagents that probe these unexplored parts of structure space. A weakness of this approach is that it does not allow the simultaneous optimization of all R groups.

For cases when an X-ray structure of the target receptor is available, the ideal is to bias the design of the combinatorial library according to those products that are complementary to the binding site of the receptor. Kick, Roe and colleagues[16] report exactly this in the structure-based design of cathepsin D inhibitors. The structure of the protease cathepsin D was determined by Baldwin *et al.*[35] Next, the Ellman group designed a combinatorial synthesis plan for creating inhibitors of cathepsin D (Figure 8). With three sites of diversity and commercially available reagents, they faced the possibility of $>10^9$ compounds. Adapting the BUILDER software originally reported by Lewis *et al.*[36], Roe and Kuntz (Ref. 37 and D.C. Roe, PhD thesis, University of California, 1995) developed the module CombiBuild to select those molecules that best complement the receptor from those compounds feasible by the combinatorial scheme (Figure 9). Their results were striking: over 6% of the molecules selected were determined to have better than 1 μM affinity to cathepsin D, with the most potent compound having a K_i = 73 nM. They compared explicitly this structure-based selection procedure with a purely diversity-based one, the latter method yielding fewer than 3% of the molecules with submicromolar affinity (Table 1); at a threshold of 330 nM, the structure-based selection procedure yielded seven times as many hits as the diversity-based procedure. They were fortunate that this particular system was free of the confounding factors that often bedevil design based on protein structure; for example, the receptor did not undergo large-scale ligand-induced conformational changes[38] and they did not observe 'precipitous binding modes', where very similar molecules bind in totally different ways, as described for thrombin inhibitors by Hilpert *et al.*[39]

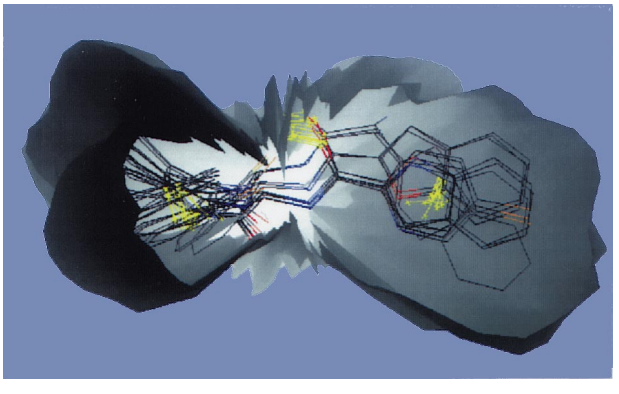

Figure 7. *Inferred shape of binding cavity for CCK antagonists*[31].

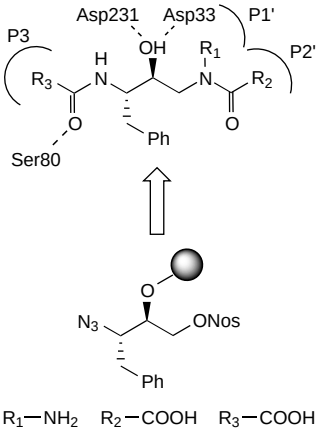

Figure 8. *Combinatorial design for cathepsin D inhibitors*[35].

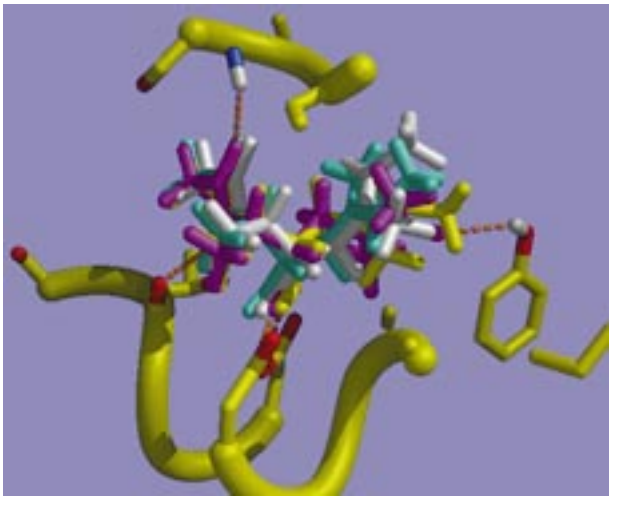

Figure 9. *Molecules selected from synthetically possible cathepsin D inhibitors, based on complementarity to X-ray structure (D.C. Roe, PhD thesis, University of California, 1995).*

Table 1. Number of compounds inhibiting cathepsin D (Ref. 16)

Inhibitor concentration	Directed library	Diverse library
100 nM	7	1
330 nM	23	3
1 μM	67	26

Other selection methods

Experimental design techniques have been explored as a means for designing sublibraries. Young and Hawkins[40] employed a statistical technique, formal inference-based recursive modelling (FIRM), to show that a 'fractional factorial design' imparts the same information that a 'full factorial design' would yield, demonstrating that among all 512 D/L combinations in a nonapeptide, only 128 peptides are needed to identify the main trends. FIRM is one example of a technique referred to as recursive partitioning, which utilizes statistical hypothesis testing to identify the single most important variable for dividing a data set into homogeneous parts. The procedure is recursive and stops when each subgrouping of the data can be divided no further.

Higgs *et al.*[41] tested a variety of different statistical design techniques in considering the problem of selecting screening sublibraries. They found that a library of 90,865 commercial compounds could be trimmed to a sublibrary of 28,896 compounds, using over 100 rules for rejecting bad structures.

The discussion so far has focused on selecting sublibraries to optimize affinity. While a potent ligand is a good milestone in the discovery of a drug, there are still many hurdles that stand in the way, many of these related to issues of absorption, distribution, metabolism and excretion (ADME). In a bold attempt to begin to address such issues, Lipinski *et al.*[42] recently described a series of simple rules to assess the ADME liabilities of a given molecule. By studying properties of drugs that are on the market, or in late-phase clinical trials, they concluded that the following rules should be observed in order to minimize ADME problems in the development of a compound:

- Molecular weight <500
- Number of H-bond donors ≤ 5
- Number of H-bond acceptors ≤ 10
- Calculated log of octonol/water partition coefficient (ClogP) <5

Lipinski's rules should be viewed as guidelines, not absolutes, yet they appear to provide satisfactory ways of biasing the design of libraries for pharmaceutical discovery. Additional aspects that might also be considered in biasing the selection procedure in library design are practical issues, such as reagent cost and availability, as has often been pointed out[14–16,43].

Open issues and controversies

Below are some of the fundamental issues for which a consensus does not exist in the literature. Each of these is an area for future work.

Should only reactant diversity be assessed?

Gillet, Willett and Bradshaw[44] have analyzed this question explicitly and conclude that assessing the diversity of the products of combinatorial syntheses leads to significantly more diverse libraries than from assessing merely the diversity of the reactants. This result appears to disagree with the observation of Patterson *et al.*[24], that 2D descriptors based only on reactants perform better than 2D descriptors computed from the products. Our own experience leads us to favor the Gillet *et al.* side of this controversy; clearly more studies are needed.

Are 3D descriptors better than 2D descriptors?

One of the most debated contributions has been that of Brown and Martin[45], in which they attempt to address this issue directly. They conclude that the 3D descriptors used in the Unity 3D database software perform significantly worse than standard 2D descriptors, for example, MACCS-keys.

By contrast, Patterson *et al.*[24] concluded that certain types of 3D descriptors performed equally well to standard 2D descriptors, while 3D descriptors composed from 'H-bonding CoMFA fields' were, overall, the best of those analyzed. There is a tendency to expect a priori that 3D descriptors should perform better, because the basis of the biological activity of a molecule is fundamentally a 3D property.

The authors' opinion on this matter is that it depends on the type of library to be designed. For screening libraries, where no reference is made to the 3D structure of a specific target, 2D descriptors appear to function reasonably well, and generally at a much smaller computational cost than most 3D descriptors. However, for the design of libraries directed to a specific target, whether an X-ray

structure is available or not, the use of 3D information becomes critical.

Can a 'representative sample' be truly representative?

All diversity-based selection procedures assume implicitly that the answer to this question is yes. However, one point that is not addressed in the diversity literature is that the answer to this question depends heavily on the structure of the chemical space. Patterson *et al.*[24] acknowledge that a fundamental assumption underpinning their work is that similar molecules will have similar biological properties. In general, our experience indicates that this is true, though much of medicinal chemistry is devoted to discovering those unusual cases where small modifications in structure lead to large changes in activity.

The problem is illustrated schematically in Figure 10. If chemistry space were two-dimensional, with the third dimension of height representing biological activity, the grid layout shown in Figure 10a would be sufficient to capture the gradually rising 'hills' of activity as the structure is changed. However, if the chemical space behaves as in 10b, where the activity hill is smaller than the grid size, then any sublibrary chosen that did not include that particular grid cell would miss an important peak of biological activity, that is, the larger library would contain a hit but the sublibrary would not. This observation relates to the sampling theorem of electrical engineering, which states that to sample a signal effectively, one must sample at least twice the highest underlying frequency of that signal.

Furthermore, there is an underlying danger in exploiting our existing notions of 'representative', especially when we use empirical rules (like those of Lipinski *et al.*[42]), and that is that these rules will ensure that tomorrow's drugs look like those in the past, and this could introduce an unnecessary and pernicious conservatism on our design perspective. For example, organometallic compounds are almost unknown as therapeutic agents, hence, our drug-like databases have almost no such representatives. Notions of representativeness based on such databases will ensure that, for better or worse, we continue to steer away from such molecules. Especially in the design of screening libraries, it is advantageous to be catholic about what are termed 'drug-like molecules'.

The authors' experience indicates that diversity-based sublibraries are, in general, not representative of the whole library, but we do not expect that. Once again, the question 'by what criteria will we judge whether our sublibrary is a good one?' arises. Our criteria are experimental ones, for example, the number of high-quality active compounds found in screening the sublibrary vs the number found in screening the entire compound collection. A reasonable expectation is that the number of active compounds in the sublibrary is proportionately smaller. Nonetheless, in spite of advances in HTS, screening sublibraries remains useful. Assaying the entire corporate compound collection is not always the best option. As genomics increases the number of targets to be screened and the size of our compound collections grows, the costs of the 'screen everything' approach must be considered.

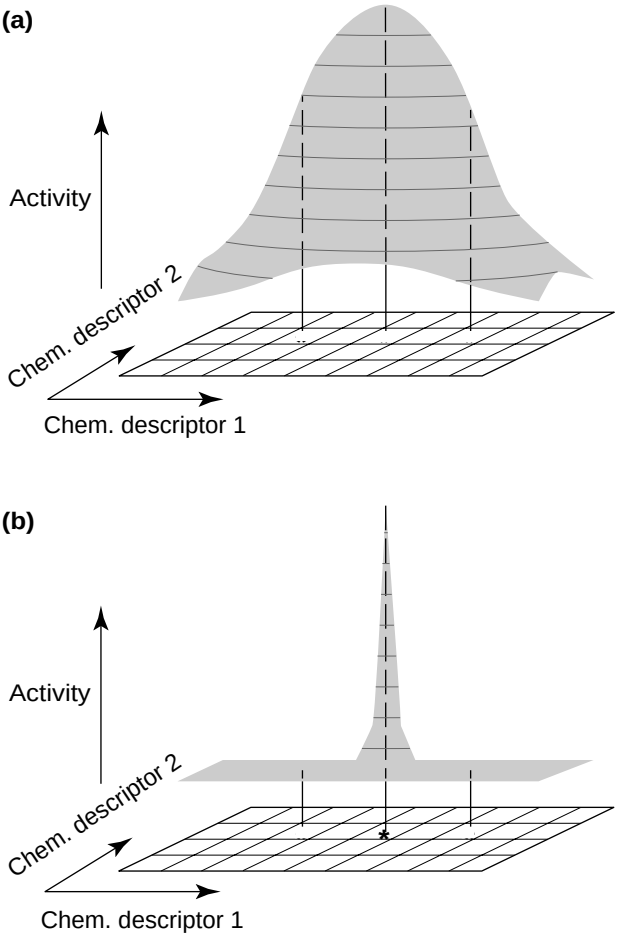

Figure 10. *How the underlying structure of chemistry space might influence any diversity-based sampling method. (a) A smoothly varying biological activity can be adequately sampled by alternating grid cells. (b) The sharply varying biological activity can only be sampled adequately by a procedure that selects the starred cell.*

Conclusions

An enormous number of creative approaches have arisen for selecting a sublibrary for screening from a larger compound collection, as well as designing combinatorial libraries. Theoretical arguments have dominated these discussions; the authors' view is that, above all, it is important to remain pragmatic – the underlying assumptions are too questionable to rely on theory alone. For the design of screening sublibraries, diversity will probably remain an important attribute to consider, but it is not the only one – molecular weight, hydrophobicity and cost are other pragmatic factors which can be considered. For the design of directed combinatorial libraries, diversity is a weak contributor to a good design, while factors related to the 3D structure of the target play a larger role.

In the choice of descriptors, 3D descriptors should not, *ipso facto*, be expected to be superior to 2D descriptors. Although some clever work has been reported in the development of 3D descriptors, more is needed.

The computer time required to perform these analyses should not be ignored. If we are unable to select sublibraries expeditiously at typical library sizes of 10^5–10^6 molecules, the experimentalists might just forego our assistance: they could either make and test all possibilities, or make their own selections based on criteria that the computational chemists would deem to be suboptimal.

Acknowledgements

We have benefited enormously from conversations with many of our colleagues at Pharmacia and Upjohn. In particular, we acknowledge R.A. Nugent, D.L. Romero, C.H. Spilman, P.K. Tomich, and especially G.M. Maggiora and our colleagues in the Computer-Aided Drug Discovery group.

We also thank Sita Nilekani for assistance with electronic literature searching; P. Willett for bringing to our attention the early work on keys from the Lynch group and other related work from that era, and emphasizing the amount of intellectual effort that went into the development of these substructural keys; and D. Roe for providing us with a draft copy of her thesis – the full thesis has only recently appeared in the public domain.

REFERENCES

1 Ellman, J.A. (1996) *Acc. Chem. Res.* 29, 132–143
2 Müller, K. (1997) *J. Mol. Struc. (Theochem)* 398–399, 467–471
3 Lewis, R.A., Mason, J.S. and McLay, I.M. (1997) *J. Chem. Inf. Comput. Sci.* 37, 599–614
4 Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*, Wiley
5 Willett, P., Winterman, V. and Bawden, D. (1986) *J. Chem. Inf. Comput. Sci.* 26, 109–118
6 Bawden, D. (1990) in *Chemical Structures 2* (Warr, W., ed.) pp. 173–176, Springer
7 Lajiness, M.S., Johnson, M.A. and Maggiora, G.M., (1989) *QSAR: Quantitative Structure–Activity Relationships in Drug Design* (Fauchère, J.L., ed.), Alan R. Liss, New York
8 Lajiness, M.S. (1990) in *Computational Chemical Graph Theory* (Rouvray, D.H., ed.), Nova Science Publishers, New York
9 Maggiora, G.M. *et al.* (1988) *Math. Comput. Modeling* 11, 630–634
10 Geysen, H.M. and Mason, T.J. (1993) *Bioorg. Med. Chem. Lett.* 3, 397–404
11 Pavia, M.R., Sawyer, T.K. and Moos, W.H. (1993) *Bioorg. Med. Chem. Lett.* 3, 387–396
12 Bunin, B.A. and Ellman, J.A. (1992) *J. Am. Chem. Soc.* 114, 10997–10998
13 DeWitt, S.H. *et al.* (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 6909–6913
14 Ferguson, A.M. *et al.* (1996) *J. Biomol. Screening* 1, 65–73
15 Garr, C.D. *et al.* (1996) *J. Biomol. Screening* 3, 179–186
16 Kick, E.K. *et al.* (1997) *Chem. Biol.* 4, 297–307
17 Adamson, G.W. *et al.* (1973) *J. Chem. Doc.* 13, 153–157
18 Adamson, G.W. *et al.* (1974) *J. Chem. Doc.* 14, 44–48.
19 Hodes, L.J. (1976) *J. Chem. Inf. Comput. Sci.* 16, 88–93
20 Hodes, L.J. and Feldman, A. (1978) *J. Chem. Inf. Comput. Sci.* 18, 96–100
21 Graf, W. *et al.* (1979) *J. Chem. Inf. Comput. Sci.* 19, 51–55
22 Howe, W.J. and Hagadone, T.R. (1982) *J. Chem. Inf. Comput. Sci.* 22, 8–15
23 Cramer, R.D. *et al.* (1996) *J. Med. Chem.* 39, 3060–3069
24 Patterson, D.E. *et al.* (1996) *J. Med. Chem.* 39, 3049–3059
25 Brown, R.D. and Martin, Y.C. (1997) *J. Med. Chem.* 40, 2304–2313
26 Hassan, M. *et al.* (1996) *Mol. Diversity* 2, 64–74
27 Sadowski, J., Wagener, M. and Gasteiger, J. (1995) *Angew. Chem., Int. Ed. Engl.* 34, 2674–2677
28 Mason, J.S. and Pickett, S.D. (1997) *Perspect. Drug Des. Discovery* 7, 1–29
29 Good, A.C. and Lewis, R.A. (1997) *J. Med. Chem.* 40, 3926–3936
30 Wipf, P. *et al.* (1997) *Bioorg. Med. Chem.* 5, 165–177
31 Van Drie, J.H. and Nugent, R.A. (1998) *SAR and QSAR Environ. Res.* 9, 1–21
32 Van Drie, J.H. (1996) *J. Comput.-Aided Mol. Design* 10, 623–630
33 Van Drie, J.H. (1997) *J. Comput.-Aided Mol. Design* 11, 39–52
34 Van Drie, J.H. (1997) *J. Chem. Inf. Comput. Sci.* 37, 38–42
35 Baldwin, E.T. *et al.* (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 6796–6800
36 Lewis, R.A. *et al.* (1992) *J. Mol. Graph.* 10, 66–78
37 Roe, D.C. and Kuntz, I.D. (1995) *J. Comput.-Aided Mol. Design* 9, 269–282
38 Rockwell, A. *et al.* (1996) *J. Am. Chem. Soc.* 118, 10337–10338
39 Hilpert, K. *et al.* (1994) *J. Med. Chem.* 37, 3889–3901
40 Young, S.S. and Hawkins, D.M. (1995) *J. Med. Chem.* 38, 2784–2788
41 Higgs, R.E. *et al.* (1997) *J. Chem. Inf. Comput. Sci.* 37, 861–870
42 Lipinski, C.A. *et al.* (1997) *Adv. Drug Deliv. Rev.* 23, 3–25
43 Martin, E.J. and Critchlow, R.E. (1997) *213th ACS National Meeting*, 13–17 April, San Francisco, CA, USA (CINF 0003)
44 Gillet, V.J., Willett, P. and Bradshaw, J. (1997) *J. Chem. Inf. Comput. Sci.* 37, 731–740
45 Brown, R.D. and Martin, Y.C. (1996) *J. Chem. Inf. Comput. Sci.* 36, 572–584