The *in silico* world of virtual libraries

Andrew R. Leach and Michael M. Hann

Combinatorial chemistry has provided medicinal chemists with an unprecedented ability to synthesize large numbers of molecules. However, early experience was that this did not result in any increase in the number of real candidates for lead optimization. Attention has increasingly focussed on the use of computational techniques for the design of combinatorial libraries. This review will describe some of the issues that have been considered in this area and discuss some of the possible developments in the near future.

Many large pharmaceutical companies started to make significant investments in combinatorial chemistry and HTS in the early 1990s. In some cases, specialized start-up companies were purchased, such as the acquisition of the Affymax Research Institute by GlaxoWellcome in 1995 or of CombiChem by DuPont Merck in 1999. In other cases, multimillion-dollar alliances have been established, such as that between Neurogen and Pfizer, also announced in 1999.

In the early days of combinatorial chemistry, much of the emphasis was on the experimental technology, to the extent that just to make a library was often considered a major achievement. Solid-phase synthesis methods were new to many medicinal chemists as was the use of automation. By the mid-1990s, some assessment could be made of their utility in providing new lead series. Generally, the results were disappointing, with the hit-rate often being very low (even zero!), and significantly lower than was achieved using compounds derived from 'historical' collections of individual molecules*. Some of the reasons for this were undoubtedly technological (e.g. sample purity) but it soon became clear that simply relying on numbers alone was not going to provide drug discovery with the tool that it sought to boost its productivity. Although combinatorial chemistry makes it possible to synthesize molecules at lower unit cost than traditional synthetic methods, this is of little value if the compounds do not have any useful biological activity.

This article will describe some of the main problems in library design over the past 2–3 years. These will be considered in approximately historical order and will be based to some extent on the experiences at GlaxoWellcome in the UK. It is not the authors' intention to provide an in-depth review of the field; for this, the interested reader is referred to recently published reviews[1–5].

'Diversity' libraries

In the early days of chemical libraries, much of the computational effort was focussed on the development of registration systems that could support the synthesis, HTS and re-synthesis processes. However, it also became clear relatively quickly that computational methods had an important role to play in the design of the libraries prior to synthesis, as it is impossible to make every molecule that could be made. There are many reasons why this is so, ranging from the availability of the starting materials to unforeseen synthetic problems. One very potent argument is the expense: although the unit cost-per-molecule is typically lower for combinatorial chemistry than for the

* A hit is a compound of known structure that shows a dose–response in a primary screening assay; a lead series is a compound series that has sufficient potential (based on various factors such as potency, selectivity, novelty) to be considered as a suitable chemical starting point for lead optimization.

*Andrew R. Leach and Michael M. Hann, Computational Chemistry and Informatics Unit, Medicines Research Centre, GlaxoWellcome Research and Development, Gunnels Wood Road, Stevenage, Hertfordshire, UK SG1 2NY. *tel: +44 1438 763383, fax: +44 1438 764918, e-mail: arl22958@glaxowellcome.co.uk

 1359-6446/00/$ – see front matter PII: S1359-6446(00)01516-6

traditional, 'one-at-a-time' approach (though by no means is this always true), the overall cost of making a library can be very significant. It is therefore necessary to select from the large virtual world of molecules that could be made, those that are appropriate to make; computational methods are ideally suited to this virtual or *in silico* pre-screening.

The main focus at this time was on producing libraries that were as diverse as possible. This of course begs the question, 'What is meant by diversity and how do you measure it?' Many groups have attempted to answer this question and, in the process, have developed some interesting algorithms and software[6]. What is common to these is the use of descriptors to characterize and differentiate molecules. There are numerous possible descriptors for a molecule[7,8]. Some of these could be experimentally measured values (such as solubility), and others might be calculated (such as a polar surface area). By definition, a unique set of descriptors can be found for every distinct compound. In library design, calculated descriptors are invariably used because this enables decisions to be made before molecules are actually produced. Moreover, as the number of compounds in a virtual library can be very large, it is often necessary to limit the process to those descriptors that can be computed rapidly. Until recently, this usually meant relying on so-called two-dimensional (2D) descriptors (calculated from the connection table, the computer equivalent of the molecule's chemical diagram). It is now also sometimes possible to use descriptors derived from a consideration of the molecule's 3D structure.

The key question is to decide which descriptors are important. Some descriptors are probably irrelevant to actual biological assays (e.g. the cost of a compound) but are important for other reasons. Of the remainder, some descriptors will be relevant to some targets and not to others. Finding the optimal set of descriptors for each target is a time-consuming process, which explains the attraction of those descriptors that claim to be universally relevant. Some of the issues surrounding the meaning of diversity will be discussed further at the end of this article.

There is now a distinct trend away from the synthesis of diversity libraries and towards libraries that focus on one specific biological target or a group of related targets. However, the tremendous power of combinatorial chemistry should still be utilized, to rapidly and efficiently explore chemical space. The main aim now is to achieve the correct balance between diversity and focus. There are no firm rules for achieving this; how best to realize this balance will depend on many factors such as restrictions of the chemical synthesis, availability of monomers, timescales and assay capacity. However, a good general rule is that the degree of diversity required is inversely related to the quantity of knowledge known about the target(s) for which those libraries are designed[9].

Drug-like libraries

Another complex notion that has been investigated is that of drug-likeness[10,11] There are many qualities that confer activity against some relevant biological target but consideration should also be given to the pharmacological and toxicological properties (in addition to issues such as cost of goods) required to make the molecule a drug candidate. Computational chemists have striven to capture the essence of what makes a molecule a drug rather than just any ordinary organic molecule, and they have had some success. However, it is important to realize that these computational tools can never fully capture the complete essence of this complex problem.

Perhaps the simplest way to apply some measure of selection or focus involves the use of substructural filters to remove molecules containing certain undesirable groups (e.g. alkyl halides, acid chlorides) that will invariably give a positive reading in any biological assay owing to factors such as their reactivity towards proteins or tendency to decompose[12]. It is also important to consider other properties of the molecules. For example, early libraries frequently contained highly flexible molecules. When a flexible molecule binds to a protein to form a bimolecular complex, it will have a larger entropic penalty to pay than a less flexible molecule (all other factors being equal) because of the reduction in the conformational degrees of freedom. As Fig. 1 shows, there can be considerable variation in the distribution of the number of notable bonds in various libraries, with some libraries being quite similar to the drug-like collections but others being very different. Such properties can be calculated easily and rapidly from the molecule's 2D-structure and so can be applied to large virtual libraries. Moreover, properties such as these can be used as the input to a computational model that predicts if the molecule is drug-like. The model might be rather simple, as is the case of the Lipinski's 'rule of 5' for oral bioavailability[13]. Alternatively, a more detailed approach is taken in which a neural network[14,15] or a genetic algorithm[16] is used to construct the computational model. In this case, the algorithm is provided with sets of drug-like and non-drug-like molecules and it aims to derive a model that best discriminates between them.

It is also possible to extend this type of analysis to generate models that can discriminate certain classes of compounds (e.g. those likely to be active at seven-transmembrane receptors) and to model so-called ADME (absorption,

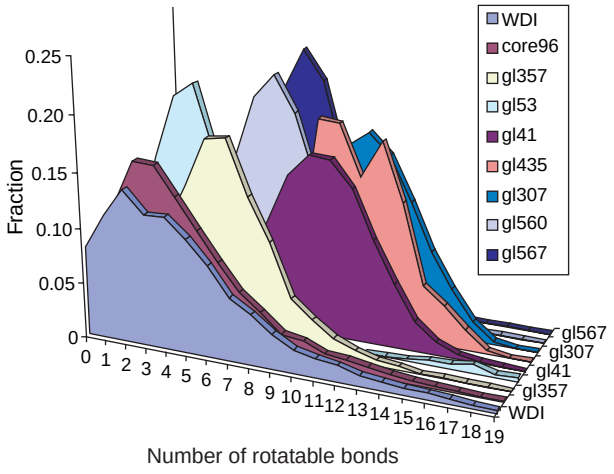

Figure 1. *The number of rotatable bonds for various early combinatorial libraries, the World Drug Index (WDI) and a representative set of GlaxoWellcome 'historical' compounds (labelled core96). The y-axis indicates the fraction of molecules in the set that contains a given number of rotatable bonds. As can be seen, some libraries (e.g. gl53) have a distribution very similar to that of the WDI or core96, while other libraries (e.g. gl307, gl435, gl41) contain much more flexible molecules.*

distribution, metabolism and excretion) properties. The important feature of these methods is that they involve the calculation of relevant properties from the structure, which are then used to assess the quality of that particular molecule. It should, however, always be remembered that such approaches are only general in nature and that any specific target might require molecules that violate one or more of these general criteria. Indeed, three of the top-selling GlaxoWellcome compounds would probably be considered inappropriate by many of the drug-likeness filters currently implemented [ranitidine, which contains a nitro-group; AZT (azidothymidine), which contains an azido-group; and salmeterol, which has an extremely long hydrocarbon chain that increases both the partition coefficient and the flexibility].

To compute the properties of a virtual library, the molecular structures must obviously be in some computer-readable format. The size of most virtual libraries means that it is not a practical proposition to consider drawing these structures by hand! It is therefore necessary to enumerate the virtual library.

Library enumeration

The term enumeration, when applied to a combinatorial library, refers to the generation of the connection tables for the product structures in a real or virtual library. There are generally two different approaches to the enumeration problem. The first of these is often referred to as 'fragment marking'. In this method, a central core template, common to all product structures, is identified. The template will contain one or more points of variation where different substituents (R-groups) can be placed. By varying the R-groups at the points of substitution, different product structures can be generated. To enumerate a combinatorial library, it is first necessary to construct sets of R-group substituents from the relevant monomer sets. Enumeration of the full library corresponds to systematic generation of all possible combinations of R-group substituents at the different points of variation.

The alternative approach is to use the computational equivalent of a chemical reaction, or 'reaction transform'. Here, there is no need to define a common template nor to generate sets of 'clipped' reagents. Rather, the library can be enumerated using the initial reagent structures as input and the chemical transforms required to operate on them. This more closely replicates the stages involved in the actual synthesis, in which reagents should react together according to the rules of synthetic chemistry.

The key elements of the fragment-marking and transform approach can be illustrated using, as an example,

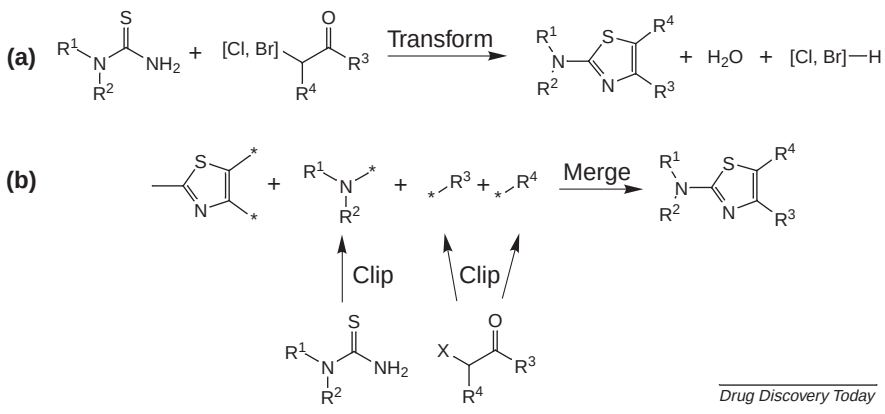

Figure 2. *Comparison of (a) the reaction transform and (b) the fragment marking approaches to library enumeration for the synthesis of aminothiazoles.*

the synthesis of aminothiazoles from thioureas and α-halo ketones (Fig. 2)[17]. With the reaction transform approach (Fig. 2a), an appropriate transform would be defined. The enumeration engine then applies this transform to the initial starting materials (i.e. the thiourea and the α-halo ketone) to produce the aminothiazole (with water and the hydrogen halide as byproducts). Using the fragment-marking approach (Fig. 2b), three sets of 'clipped' fragments would be constructed (two from each α-halo ketone and one from each thiourea), which would then be grafted on to give the central thiazole core to produce the appropriate products.

Advantages and disadvantages of fragment-marking and transform approaches to enumeration

In favour of the mark-up approach is the fact that for some library types (i.e. those that most obviously fit the 'core plus R-group' definition), this method can be the fastest way to enumerate the library. This is because the fragment-marking approach only involves some rather elementary connection table operations after generation of the R-groups. Although most systems offer automated methods of generating the R-groups (i.e. clipping algorithms), problems invariably arise that need to be corrected by hand. This can make the fragment-marking approach time-consuming to perform for a non-expert, unless sets of predefined R-groups are already available. In addition, there are certain reactions that are not properly handled by the basic fragment-marking approach, one well-known example being the Diels–Alder reaction, where this approach would generate several extraneous and incorrect products. Moreover, in some cases, there is no clear core structure (e.g. oligomeric libraries such as peptoids). The advantages of the reaction transform method include the ability to enumerate directly from the reagents without having to perform any pre-processing, and to re-use the same transforms many times (after they have been defined). However, this method requires more computational steps and so is typically slower. Perhaps the key advantage is that this approach models the actual chemical steps involved in the experiment, so bringing the experimental and computational systems closer together.

Combinatorial subset selection

For any given synthetic scheme, the main issue in combinatorial library design is monomer selection, the objective of which is to identify those monomers (reagents) that, when combined together, provide the optimal combinatorial library. Here, 'optimal' refers to the library that best meets the prescribed objectives: it might be the most diverse, have the maximum number of molecules that could fit a 3D pharmacophore or a protein binding site, be the best match to a particular distribution of some physicochemical property, or be some combination of these or other criteria. An important consideration when designing a combinatorial library is the subset selection constraint. In a 'true' combinatorial library of the form A×B×C, every molecule from the set of reagents A reacts with every molecule from B and every molecule from C to generate $n_A \times n_B \times n_C$ product structures, where n_A, n_B and n_C are the numbers of reagent molecules in A, B and C. Typically, there will be many more possible reagents in A, B and C available than can actually be incorporated into the library, hence the need to select the subset of monomers that give rise to the optimal library. Given that the number of possible reagents A is N_A, etc., the size of the virtual library is $N_A \times N_B \times N_C$. The number of ways of selecting n objects from N is ${}^{N}C_{n}$, and therefore the number of different combinatorial libraries of size $n_A \times n_B \times n_C$ that could be made for this three-component library is ${}^{NA}C_{nA} \times {}^{NB}C_{nB} \times {}^{NC}C_{nC}$. If there are 100 possible reagents for each of A, B and C and the aim is to make a 10×10×10 library, then the number of different libraries that could be made is approximately 10^{40}. Identifying the one optimal library from this extremely large number of possible libraries is clearly a difficult problem that cannot be solved by a systematic examination of every possible solution.

Tackling the subset selection problem

Computational methods such as genetic algorithms[18,19] have been used to tackle the subset selection problem. These methods evolve possible solutions until either no better solution can be found or until the pre-determined number of iterations is exceeded. The real advantage of these more complex methods is their ability to simultaneously optimize many different factors. An example is the program SELECT, jointly developed by Val Gillet and Peter Willett (Sheffield University, Sheffield, UK) and John Bradshaw and Darren Green (GlaxoWellcome, Stevenage, UK)[20]. With SELECT, a user can design a library so that it has a distribution of properties such as the MW and/or a calculated partition coefficient similar to a target set of molecules (such as the World Drug Index). Another example is the GALOPED algorithm devised by Brown and Martin (Abbott Laboratories, Chicago, IL, USA), which can optimize the diversity of libraries while minimizing the effort required to deconvolute the biological hits by MS techniques[21]. A program such as SELECT can deal with virtual libraries containing more than one million product structures.

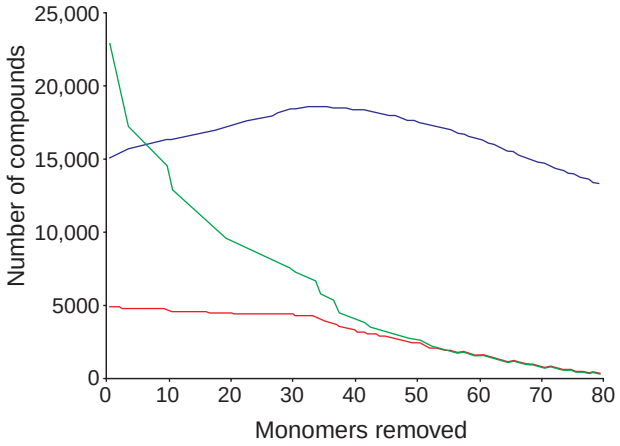

Figure 3. *Optimization of combinatorial libraries using PLUMS (GlaxoWellcome, Stevenage, UK). PLUMS successively removes monomers from the virtual library. The number of hits (red) indicates the number of virtual products that are 'acceptable' (e.g. match a 3D pharmacophore; have properties within a given range). The library size (green) indicates the total number of molecules in the virtual library. The efficiency/efficacy score (blue) provides a balance between these two measures. As can be seen, the library size typically drops more significantly during the first few iterations than does the number of hits. The maximum in the score indicates the 'optimal' library size.*

Some simpler methods are also available for tackling the subset selection problem. These alternative methods typically use a less exhaustive search of the space and therefore are faster. One such approach is Monomer Frequency Analysis (University of North Carolina, Chapel Hill, NC, USA)[22] in which the frequencies with which monomers appear in 'acceptable' product molecules are computed and compared with the random frequency value. Only monomers whose frequencies are above this threshold are considered for the experiment. An alternative method called PLUMS[†], invented by Gianpaolo Bravi (GlaxoWellcome), tackles the problem by successively removing monomers from the virtual library. At each stage, the monomer chosen for removal is the one that adds least value to the library. An optimization function guides the algorithm; this function comprises two terms, effectiveness and efficiency. The effectiveness of a virtual library is the number of molecules contained within it that meet the desired criteria (such as having certain properties within predefined ranges, or that fit to a 3D pharmacophore). The efficiency is the ratio of the number of such molecules to the total size of the library. It is usually found that the initial virtual library is highly effective (i.e. contains many favourable molecules) but is rather inefficient. As monomers are removed, the efficiency improves but the effectiveness falls. The 'best' library is often the one with a sensible balance between effectiveness and efficiency. This can be monitored using a function that is the simple sum of the two terms (Fig. 3).

This type of library design is often known as product-based monomer selection as it is the properties of the product molecules that determine the ultimate monomer selections. Enumeration is clearly key to this approach as it requires product structures to be generated. The alternative is monomer-based selection, where only the properties of the individual monomers are considered and not the properties of the product molecules. The main advantage of monomer-based selection is that the size of the search space is much smaller; product-based selection requires direct or indirect consideration of the $N_A \times N_B \times N_C$ potential product molecules whereas in monomer-based selection, only the $N_A + N_B + N_C$ monomers need be considered. It has been shown that product-based selection produces superior results (as would be expected)[23], but it is also clear that, in some cases, the virtual libraries will be so large that full enumeration might be impossible, and some combination of the two approaches could be required. Moreover, more recent studies appear to suggest that under some circumstances, a monomer-based selection might be comparable in quality to a product-based selection, depending on the descriptors used to characterize the library[24]. In addition, some experimental methods such as structure–activity relationships (SAR)-by-NMR (Ref. 25) are ideally suited to working in monomer space.

Monomers for combinatorial libraries

The quality and success of a library is crucially dependent on the monomers used in its synthesis. This is particularly true in the hits-to-leads phase, where rapid exploration might be required of the chemical space around a hit from a biological screen. Despite the apparently large numbers of potential monomers available commercially [for example, the current version of the Available Chemicals Database (ACD; MDL Information Systems, San Leandro, CA, USA) contains >20,000 carboxylic acids], many of these are unsuitable for library synthesis. There might be many reasons for this: some contain functionalities that are incompatible with the chemical synthesis, others do not

[†] Bravi, G. *et al.* PLUMS: a program for the rapid optimization of focussed libraries. *5th International Conference on Chemical Structures*, 6–10 June 1999, Noordwijkerhout, The Netherlands, Abstract 7.

have the appropriate protecting group functionality, and others are too large, too flexible or contain groups that would not be compatible with drug-like molecules. Of course, there are many other potential sources of monomers than those in the ACD and many companies have on-going programmes to construct databases of molecules available for acquisition. There might also be some circumstances in which it is necessary to contemplate having monomers made to contract. Such bespoke monomers can be expensive and so it is important that a rational approach is used to decide which molecules to target.

One approach is to examine collections of biologically active molecules to assess whether it is possible to identify certain key groups or molecular fragments. Most medicinal chemists would be able to write down some of the more obvious examples such as the benzodiazepine skeleton or the penicillin ring system. However, several computational techniques have also been developed for the automatic analysis of large collections of compounds. In a series of papers, a group at Vertex (Boston, MA, USA) has examined collections of drug molecules for molecular frameworks[26] and side chains[27]. This work suggests that the number of such molecular fragments might be rather limited.

RECAP (Retrosynthetic Combinatorial Analysis Procedure; GlaxoWellcome) applies retrosynthetic rules to molecules with a particular type of biological activity[28]. This procedure can identify so-called 'privileged' monomers, which can then be used to design libraries for that class of biological target. An illustration of the application of RECAP to acebutolol, a β-adrenoceptor antagonist, is shown in Fig. 4. A relatively small number of generic bond cleavage sites are sufficient to cover a wide range of possibilities. If the process is applied to a reasonably large number of molecules with some common biological activity (typically a few thousand), then monomers can be identified that might be associated with that type of activity.

An approach termed Gridding and Partitioning (GaP; GlaxoWellcome)[‡] has also been developed. The main objective was to provide a biologically relevant descriptor space within which monomers could be compared rationally to help identify possible candidates for synthesis. The main feature of GaP is the use of a pharmacophoric description of molecules that takes conformational flexibility into account. It therefore provides a space within which candidate monomers (or sets of them) can be compared with monomers that are already available (Fig. 5). The first

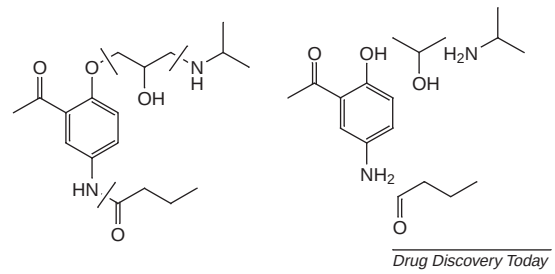

Figure 4. *Application of RECAP (Retrosynthetic Combinatorial Analysis Procedure, GlaxoWellcome, Stevenage, UK) to acebutolol. In this case, three bonds are cleaved, as shown. Note that bonds involving small substituents (e.g. the isopropyl amine) are left intact.*

step is to orientate each reagent into a common frame of reference by positioning the reagent's attachment group at the origin and aligning the reagent attachment bond along the x-axis. A systematic conformational analysis is then performed, enabling free rotation about the attachment bond. This is undertaken within a regular grid (typically 1Å cube-length). An equivalent grid is used for each pharmacophore atom type (donor, acceptor, acid, base, aromatic, donor/acceptor and heavy atom). During the conformational analysis, the positions of each of the relevant pharmacophore types are tracked, marking the appropriate cubes.

This analysis is performed not only for those monomers of potential interest but also for those that are already available in-house or are deemed readily accessible from external sources. The GaP approach provides a naturally partitioned 3D pharmacophore space. To assess a candidate monomer, those cubes that are occupied by the available monomers must first be identified. A simple 'score' for each new monomer is then given by the number of new cubes it can access. More sophisticated schemes are also possible based on the occupancy of the cells or on other properties of the molecules. The GaP method therefore provides a rational method of assessing potential candidate monomers for contract synthesis. It is important to recognize that a significant level of expert chemical knowledge is still an important part of the assessment; factors such as the ease of synthesis of a particular monomer are more difficult to incorporate into the computational scheme.

End-user software tools

In the early days, combinatorial synthesis was typically the 'province' of a small number of specialist chemists. Library design was performed by an expert computational chemist who often had to cope with many computer systems and

[‡] Leach, A.R. *et al*. Where are the GaPs? A rational approach to monomer selection. *218th ACS National Meeting,* 22–26 August 1999, New Orleans, LA, USA, Abstract CINF-002.

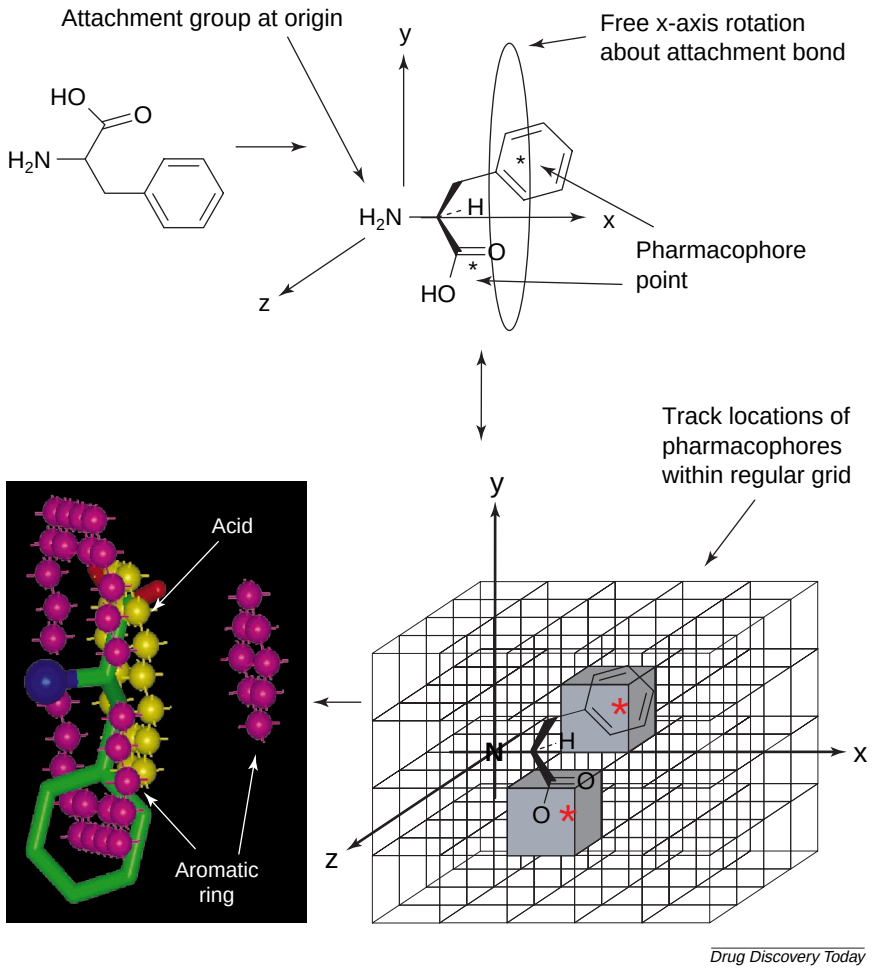

Figure 5. *The Gridding and Partitioning (GaP; GlaxoWellcome, Stevenage, UK) procedure as applied to monomers, exemplified using phenylalanine as a potential primary amine. This molecule thus contains two pharmacophoric groups (the aromatic ring and the carboxylic acid). During the conformational analysis the locations of these pharmacophoric groups are tracked within a regular grid.*

various types of software, much of which was not user-friendly. Combinatorial chemistry and parallel synthesis methods are now much more widely used in all stages of drug discovery and are not restricted to 'experts'. Likewise, software is required that can be used by bench chemists to assist in the design of their libraries. Both pharmaceutical and commercial software companies have recognized this and have developed systems accordingly. These include systems at Novartis (Basel, Switzerland; cyclops)[29], Vertex and Abbott, together with the Diversity Explorer (Molecular Simulations, San Diego, CA, USA). At GlaxoWellcome, we have developed an in-house web-based system called ADEPT (A Daylight Enumeration and Profiling Tool), which enables a bench chemist to perform the basic stages involved in library design and monomer selection[30]. The key steps available in this web-based system are shown in Fig. 6. Facilities are provided to enable chemists to perform substructure searches of databases of in-house and commercially available monomers. This initial monomer pool might be very large in the case of certain common functional groups (e.g. carboxylic acids, amines). However, the size of this initial pool is often significantly reduced in subsequent steps when the chemist eliminates from consideration those molecules containing functionality that is incompatible with their reaction scheme. It is also possible to apply some simple MW and flexibility filters at this stage.

Having identified pools of potential reagents, the virtual library can then be enumerated using the transform-based method developed by Daylight Chemical Information Systems (Santa Fe, NM, USA). This uses a language called SMIRKS[§] to define the reaction transform. SMIRKS is an extension of Daylight's SMILES (Simplified Molecular Input Line Entry System)[31] and SMARTS[¶] languages for representing molecules and molecular patterns (substructures), respectively. An example SMIRKS for the ubiquitous amide-forming reaction for a primary aliphatic amine and a carboxylic acid is as follows:

[O:0]=[C:1][O:2][H:99].[$([NH2][CX4]):3][H:100]>>[O:0]=[C:1][N:3].[H:99][O:2][H:100]

Of course, medicinal chemists are not expected to define new reaction schemes by writing SMIRKS! Rather, the user can draw their reaction in a common package such as IsisDraw (MDL Information Systems), and then copy-and-paste it into

[§] *Daylight Theory Manual* (Ch. 7), Daylight Chemical Information Systems, Santa Fe, NM, USA; http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html

[¶] *Daylight Theory Manual* (Ch. 4), Daylight Chemical Information Systems, Santa Fe, NM, USA; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

the web page where it is converted automatically by in-house software to the corresponding SMIRKS expression. However, most of the chemistries that are used in combinatorial library enumeration are covered by the transforms included by default within ADEPT.

Various properties can then be calculated within ADEPT for the enumerated library; currently, these are typically restricted to those properties that can be computed rapidly from the 2D chemical structure. The user can then refine their virtual library using these computed properties. In some cases, this will be as simple as the 'rule of 5'; in other cases, the project might have much more precise criteria. It is then necessary to identify those monomers that lead to the optimal combinatorial library. This can be achieved using PLUMS, at the speed that makes it highly appropriate for the web environment. ADEPT also has many other useful facilities for compound selection and library design. Screen-shots from ADEPT showing how a simple two-component library can be specified and the histogram output for various properties are shown in Fig. 7.

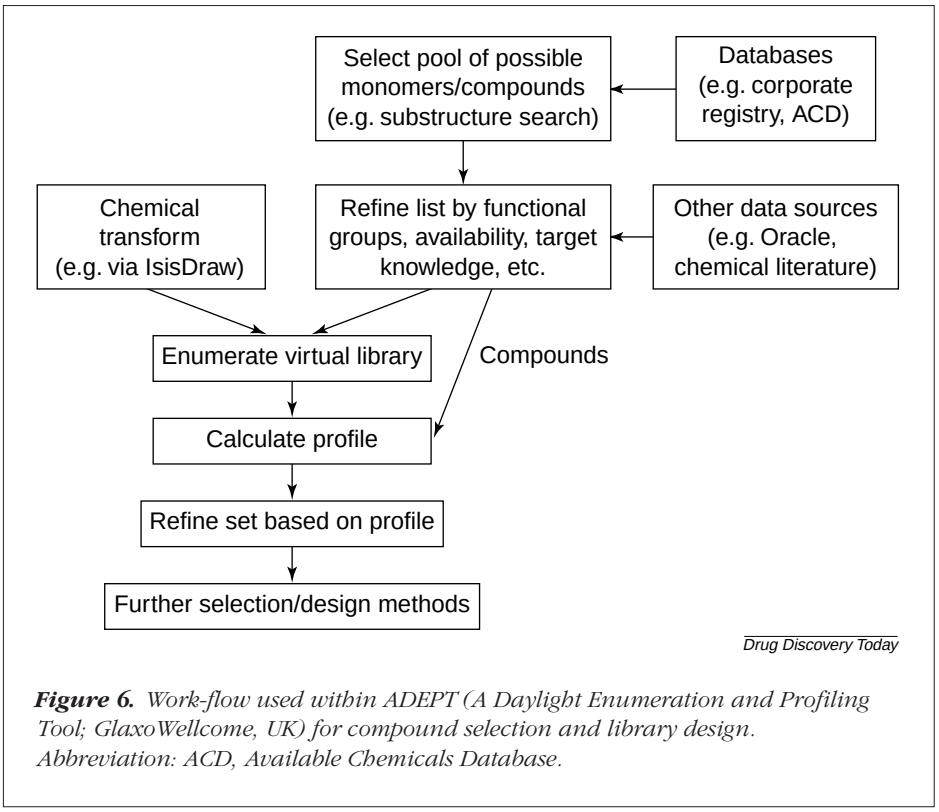

Figure 6. *Work-flow used within ADEPT (A Daylight Enumeration and Profiling Tool; GlaxoWellcome, UK) for compound selection and library design. Abbreviation: ACD, Available Chemicals Database.*

Taking 3D properties into account

Much of the early library work involved working with 2D representations of molecular structure. For large problems (involving libraries containing many millions of structures), this is still the case. However, it is the 3D properties of molecules that ultimately determine their behaviour. There are several ways in which these 3D aspects can be taken into account. Perhaps the most obvious case arises when the structure of the biological target is available (e.g. from X-ray crystallography), and many groups have developed software tools for such structure-based library design[32–36,**]. Alternatively, a 3D pharmacophore derived from a series of active molecules might be available. It is more time-consuming to consider the 3D properties of molecules, and so access to appropriate computational resources can become a major factor, at least with current algorithms. Computer 'farms', built using cheap PC chips running the Linux operating system might offer one possible solution. It is also important to ensure that only those molecules that are truly of interest are subjected to the more time-consuming 3D computational analysis. For example, the analysis could be restricted to those chemistries that have been 'tried and tested' in-house and those monomers that have been shown to work in these chemistries. This is a strategy that GlaxoWellcome has adopted when building large 3D databases of virtual libraries (and involves consideration of conformational flexibility for several million structures). This therefore provides confidence that any library selected using these databases could rapidly be synthesized with a high probability of success in a given screening campaign.

The future

What does the future hold? Combinatorial chemistry has developed so rapidly over the past few years that it is very difficult to make any kind of prediction. However, it is clear that there are some key issues in the way that combinatorial chemistry is practically applied in drug discovery and the role of computational methods in supporting that process. Many of these issues are perhaps more concerned with practice than theory, such as the need for adequate

** Walters, W.P. *et al*. Skizmo – A new program for rapid conformational analysis and structure-based design of combinatorial libraries. *217th ACS National Meeting,* 21–25 March 1999, Anaheim, CA, USA, Abstract COMP-012.

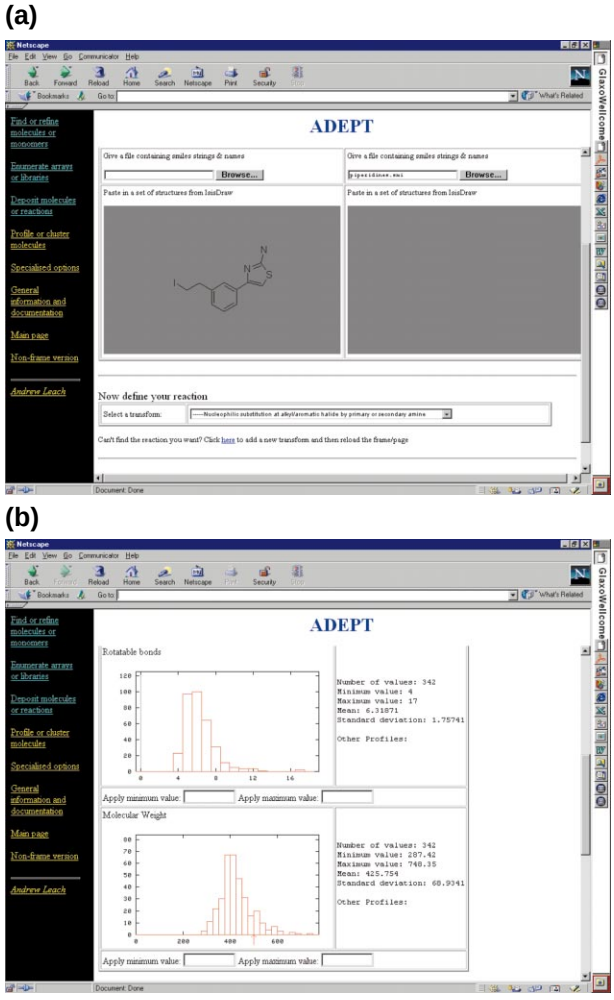

Figure 7. *Screen shots from ADEPT: (a) shows how a simple two-component library comprising an aminothiazole template and a series of piperidines[41] would be specified within ADEPT; (b) shows how the distribution of rotatable bonds and MW can be displayed in histogram format for the enumerated virtual library.*

supplies of appropriate monomers, the development of new solid-phase chemistries and the alignment of synthesis and screening resources. However, there are clearly some aspects of library design and synthesis where the development of new computational methods (or the re-discovery of old ones) might reasonably be expected. One such aspect is in predicting the likely reactivity of combinations of monomers[37]. Another is the use of 3D information in the form of a protein structure or a 3D pharmacophore. Currently, only relatively modest numbers of molecules can be assessed in this way and new algorithms might be required for significantly larger virtual libraries. Markush representations, originally developed for the computer representation of patents, have tremendous potential for the enumeration of large libraries, for the calculation of properties and for the comparison of libraries, both real and virtual[38]. These techniques, when combined with sophisticated optimization procedures, such as the next generation of genetic algorithms, will enable the exploration of more of the virtual chemical space.

We would like to conclude with some thoughts on chemical 'diversity'. This has been a widely used (and abused) term in the past 5–10 years in drug discovery. We believe that as far as numbers of molecules are concerned, there is really no appropriate definition of the term. This is because of the vast number of molecules that could be considered candidates for synthesis, acquisition or extraction. It has been estimated that there are more than 10^{18} potential drug-like molecules that could be synthesized for screening[39]. Such numbers are well beyond the capacity of current computers, and they almost defy human comprehension. While the precise scale of the problem can be debated, it is obvious that the 10^7–10^8 molecules currently known is only a 'drop in the ocean' of what is possible (i.e. a maximum sampling rate of approximately 1 in 10^{12}). To practically get to grips with the problem of chemical diversity requires some thought as to how far into the vast chemical universe to go and in what direction.

One view of diversity that has been widely used is to consider it as the opposite of chemical similarity. Although this makes the problem more easily understood, it is also fraught with difficulties. This is because different biological systems respond differently when presented with chemically 'similar' molecules and so there is no universal similarity measure that provides for our needs. In some cases, adding a methyl group can abolish activity whereas in other molecules, it makes no difference or even enhances activity – no universal score can cope with this in the absence of a true and predictive understanding of the target. Only in a few cases can the effect of such small structural changes be confidently predicted *a priori* (usually when detailed X-ray crystallographic information is available). Combinatorial chemistry provides the means to explore more of the chemical space than was traditionally possible and therefore, in some ways, enables us to mask our inability to make accurate predictions. However, the vast size of chemical space means that choices must be made concerning which particular region(s) of the chemical space to concentrate on and what implications this has for the type of molecules that should be made and the type of assays that should be run. Given that it is highly unlikely

that the 'development candidate' will be found on the first attempt, it is perhaps more appropriate to expect that molecules with only modest affinity and other properties (e.g. ADME) will be found in the early stages. It is therefore important to make optimal use of the information and knowledge that is gained from these assays in the second and subsequent iterations of the process[40]. This is where computational approaches are likely to have the most impact, rather than exploring the 'true' meaning of diversity.

Acknowledgements

We are very grateful to colleagues both past and present, together with our academic collaborators for their contributions to the work described in this article. Some of these have been identified by name in the text. However, we would particularly like to recognize the input of our colleagues John Bradshaw (now with Daylight Chemical Information Systems) and Darren Green together with Peter Willett and Valerie Gillet (both at Sheffield University), all of whom have contributed to a number of the concepts described.

REFERENCES

1 Drewry, D.H. and Young, S.S. (1999) Approaches to the design of combinatorial libraries. *Chemometr. Intell. Lab. Syst.* 48, 1–20

2 Agrafiotis, D.K. *et al.* (1999) Advances in diversity profiling and combinatorial series design. *Annu. Rep. Comb. Chem. Mol. Diversity* 2, 71–92

3 Agrafiotis, D.K. *et al.* (1999) Advances in diversity profiling and combinatorial series design. *Mol. Diversity* 4, 1–22

4 Spellmeyer, D.C. and Grootenhuis, P.D.J. (1999) Recent developments in molecular diversity. Computational approaches to combinatorial chemistry. *Annu. Rep. Med. Chem.* 34, 287–296

5 Bures, M.G. and Martin, Y.C. (1998) Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* 2, 376–380

6 Martin, Y.C. *et al.* (1998) Quantifying diversity. *Comb. Chem. Mol. Diversity Drug Discovery* 369–385

7 Brown, R.D. (1997) Descriptors for diversity analysis. *Perspect. Drug Des. Discovery* 7/8, 31–49

8 Livingstone, D. (2000) The characterisation of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* 40, 195–209

9 Hann, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

10 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387

11 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of drug-likeness. *Drug Discovery Today* 5, 49–58

12 Rishton, G.M. (1997) Reactive compounds and *in vitro* false positives in HTS. *Drug Discovery Today* 2, 382–384

13 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23(1–3), 3–25

14 Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 41, 3325–3329

15 Ajay, A. *et al.* (1998) Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* 41, 3314–3324

16 Gillet, V.J. *et al.* (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 38, 165–179

17 Bailey, N. *et al.* (1996) A convenient procedure for the solution phase preparation of 2-aminothiazole combinatorial libraries. *Bioorg. Med. Chem. Lett.* 6, 1409–1414

18 Clark, D.E. and Westhead, D.R. (1996) Evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Design* 10, 337–358

19 Judson, R. (1997) Genetic algorithms and their use in chemistry. In *Reviews in Computational Chemistry* (Vol. 10) (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 1–73, VCH Publishers

20 Gillet, V.J. *et al.* (1999) Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* 39, 169–177

21 Brown, R.D. and Martin, Y.C. (1997) Designing combinatorial library mixtures using a genetic algorithm. *J. Med. Chem.* 40, 2304–2313

22 Zheng, W. *et al.* (1998) Rational combinatorial library design. 1. Focus-2D: A new approach to the design of targeted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.* 38, 251–258

23 Gillet, V.J. *et al.* (1997) The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 37, 731–740

24 Jamois, E.A. *et al.* (2000) Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* 40, 63–70

25 Shuker, S.B. *et al.* (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534

26 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893

27 Bemis, G.W. and Murcko, M.A. (1999) Properties of known drugs. 2. Side chains. *J. Med. Chem.* 42, 5095–5099

28 Lewell, X.Q. *et al.* (1998) RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38, 511–522

29 Gobbi, A. *et al.* (1997) Developing an in-house system to support combinatorial chemistry. *Perspect. Drug Des. Discovery* 7/8, 131–158

30 Leach, A.R. *et al.* (1999) Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.* 39, 1161–1172

31 Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36

32 Leach, A.R. (1997) Structure-based selection of building blocks for array synthesis via the World Wide Web. *J. Mol. Graph. Model.* 15, 158–160

33 Sun, Y. *et al.* (1998) CombiDOCK: structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Design* 12, 597–604

34 Kubinyi, H. (1998) Combinatorial and computational approaches in structure-based drug design. *Curr. Opin. Drug Discovery Dev.* 1, 16–27

35 Bohm, H-J. *et al.* (1999) Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. *J. Comput.-Aided Mol. Design* 13, 51–56

36 Jones, G. *et al.* (1999) Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. *ACS Symp. Ser.* 719, 271–291

37 Braban, M. *et al.* (1999) Reactivity prediction models applied to the selection of novel candidate building blocks for high-throughput organic synthesis of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 39, 1119–1127

38 Downs, G.M. and Barnard, J.M. (1997) Techniques for generating descriptive fingerprints in combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 37, 59–61

39 Martin, Y.C. (1997) Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Des. Discovery* 7/8, 159–172

40 Teague, S.J. *et al.* (1999) The design of lead-like combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* 38, 3743–3748

41 Selway, C.N. and Terrett, N.K. (1996) Parallel-compound synthesis: methodology for accelerating drug discovery. *Bioorg. Med. Chem.* 4, 645–654