DDT Vol. 8, No. 15 August 2003

A fresh look at pharmaceutical screening library design

Ramaswamy Nilakantan and **David S. Nunn**, Chemical & Screening Sciences, Wyeth Research, Pearl River, NY 10965, USA;
e-mail: nilakar@wyeth.com

Pharmaceutical companies identify novel medicinal lead compounds by assaying large private compound collections. With the advent of HTS, entire corporate collections are routinely screened against dozens of biological targets, thereby depleting the compound samples rapidly. There is, therefore, a big push in the industry to augment compound collections through purchase or large-scale combinatorial synthesis.

Most laboratories have been guided by the principle of structural diversity to decide which compounds to buy or synthesize for general-purpose screening – the thought being that a diverse set of compounds can cover a large variety of biological targets. Several approaches, such as simulated annealing, genetic algorithms, Monte-Carlo modelling, partitioning, and various combinations of these, have been used to maximize diversity [1–7].

So, why do we need another method for combinatorial library design? Examination of any corporate screening collection shows that it consists of clusters of varying size as well as singletons, that is, compounds that have no structural near-neighbours. The problem with such non-uniform libraries, is that singletons or compounds in small clusters are not explored to the same extent as larger clusters. Specifically, we cannot rule out activity in any series that has small representation in a library and this is why balanced libraries need to be designed where chemical series are more or less equally represented.

Library design approach

In an earlier paper, we presented the idea of building a screening library consisting of medicinally relevant scaffolds with an approximately equal number of analogs around each scaffold [8]. An analysis of HTS data from our corporate database was presented and probability arguments were used to estimate the number of analogs required around each scaffold. A refinement of that analysis is presented here, the same 18 assays being used in the present work. Over time, the assay databases have changed slightly, but remained largely the same.

As in our earlier paper, ring-scaffolds were used to substitute for chemical series, which are a somewhat subjective concept. For the rest of the discussion, the terms 'scaffold' and 'ring-scaffold' will be used synonymously with chemical series.

Analysis of HTS data

When is a series 'active'?

In the previous study [8], any series of compounds with at least one active was considered an active series. For large clusters, this can present a problem because a single hit in a large series could be a chance event. Here, an alternative, objective, approach to the identification of 'active series' is presented. Biological assays are treated as Bernoulli trials, where the probability of a compound being active is equal to the overall hit-rate. Members of an active series, however, will have a much higher underlying hit-rate. The binomial distribution can be used to determine, for each series, whether the observed number of hits is consistent with chance. For example, if there are k hits in a series of n compounds, the cumulative probability of finding k or fewer hits can be determined using Equation 1:

$$B(p,n,k) = \sum_{k=0}^{n} \binom{n}{k} p^n (1-p)^{n-k} \qquad \text{[Eqn. 1]}$$

where n = the number of compounds in the series, k = the number of hits in the series and denotes the number of combinations of 'n' taken 'k' at a time.

The critical number, k, is found such that the probability of getting k or fewer actives by chance is ≥0.99. Series containing at least k actives are deemed active. Thus, we have an objective method to identify active series.

Hit-rate in active series

Confirmed actives were determined from 18 in-house HTS assays (as shown in [8]). The compounds were clustered into series using the ring-scaffold method [8]. The overall hit-rate in active series was determined by dividing the total number of actives in active series by the total number of compounds in active series. This is a sort of grand average across all series and all assays, and provides a working estimate for the hit-rate in active series.

In our analysis there were 132 active scaffolds in the 18 assays. The active scaffolds contained a total of 4461 compounds, of which 224 were active. Thus, we can calculate a mean hit-rate in active scaffolds as 224/4461 = 0.0502 (~5%). The smallest hit-rate among the 18 assays was 0.0256, (~2%). This gives

 S1359-6446(03)02793-4

a lower-bound estimate for the hit-rate in active series.

Number of analogs required in each series

As shown previously [8], one can use simple probability arguments to derive an equation relating the number of compounds required to the overall hit-rate, and the hit-rate in active series at a given level of significance. This is given in Equation 2:

$$s = \frac{\ln(1-P)}{\ln(1-p)} \qquad \text{[Eqn. 2]}$$

where, s = number of compounds in the sample, p = hit-rate in active series, P = probability that at least one compound in the sample is a hit (could be 0.99 or 0.95 depending on the confidence level desired)

Substituting the values of P and p into Equation 2, we get:

For $p = 0.0502$, $s = 89$ [$P = 0.99$ (99% confidence level)], and $s = 58$ [$P = 0.95$ (95% confidence level)].

For $p = 0.0256$, $s = 178$ [P = 0.99 (99% confidence level)], and $s = 115$ [P = 0.95 (95% confidence level)].

The estimates for s differ, depending on the level of confidence, P, and the different estimates for p, the hit-rate in active series. If we choose the lower-bound value, it can be seen that there is a sharp increase in s. We have chosen to use the mean-value estimate of $p = 0.0502$, thereby giving an estimate of ~89 compounds for s. Armed with our estimate of s, we now turn to the task of actually building the libraries.

Design of libraries around fixed ring-scaffolds

There are practical problems with combinatorial synthesis around fixed ring-scaffolds. Combinatorial synthesis uses existing reagents and easy chemistry in return for economies of scale. Our design requires that there is a constant ring-scaffold with acyclic substituents. However, the chemistry might be such that new ring-systems are created by the reaction. Also, available reagents could have one or more rings, or they might be acyclics. Thus, a set of reagents could give rise to products that fall into different ring-scaffold classes, which, according to our definition, means different chemical series. Therefore, a method is required, by which reagent subsets can be selected to form sets of ~100 analogs around each ring-scaffold.

VLib™, *a virtual combinatorial library generator*

VLib™ is our original system, which provides an interactive interface for chemists to enumerate and analyze virtual libraries, before actual synthesis. For each proposed monomer set, the SMILES [9] representation of each reagent is searched for specified reactive functionality, which is replaced by markup atoms encoding a specific bond to be formed. Software to do this was created using Daylight toolkits [Daylight Chemical Information Systems, 27401 Los Altos, Suite #360, Mission Viejo, CA 92691, (http://www.daylight.com/)] and an additional 'transform' engine [by Bernhard Rohde of Novartis Pharma (http://www.novartis.com)]. More than 100 common transformations are included via a menu system, and users can also specify their own. Alternatively, the chemist can choose to draw starting materials directly in a pre-marked form. Simple pre-processing decodes the markup atoms into bond information, such that enumeration is simply a matter of joining fragments with the '.' character. Analyses can be applied to the reagent sets or to the enumerated products.

Steps in generating a library

There are three major steps in generating libraries. These are outlined below.

Reagent markup To make a virtual library, it is necessary to first markup the reagents according to the type of chemistry. The markup produces a conversion of the reagent into the fragment of the final molecule that it will constitute. The markup also identifies the attachment points by pseudo-atoms labelled by X1, X2 and so on.

Hashcodes from marked fragments Once the markup is done, a structure-based hashcode is calculated for each reagent. Details of the hashcode calculation can be found in an earlier publication [10]. This hashcode is calculated not from the entire marked-up fragment, but from the fragment minus the acyclic substituents. Notice that the bond-path to the markup atom is maintained, whereas other acyclic substituents are deleted. This is because the bond-path will typically form a bridge to another ring-fragment in the final molecule and will, therefore, contribute to the ring-scaffold of the final molecule; it might also be part of a newly created ring-system. A second hashcode is also calculated for the marked-up fragments by keeping ring-systems and their branch-points. This hashcode is useful in picking reagents with varying substituent patterns on the same component scaffold. The markup and hashcode calculations are illustrated in Figure 1, using an example of amide formation from acid chlorides and amines.

Combining reagents The nature of the combinatorial reaction dictates the next step in the process. If it is a three-component reaction, there will be three reagent lists, and to make a library of ~100 compounds, sets of five reagents could be chosen from each list resulting in $5 \times 5 \times 5 = 125$-member libraries. For a two-component reaction, there would be two reagent lists, and sets of ten reagents from each list could be chosen, resulting in $10 \times 10 = 100$-member libraries. Component hashcodes are used to identify reagent sets that share the

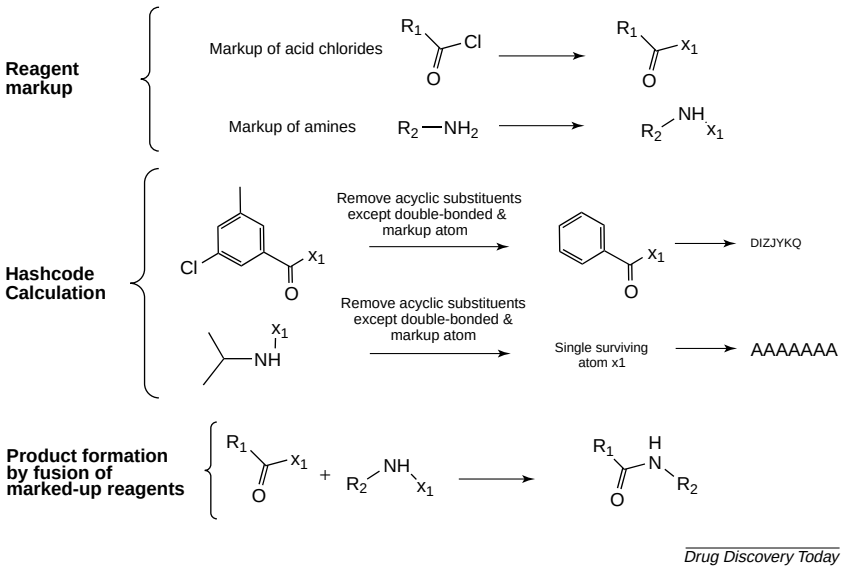


Figure 1. The markup of reagents, hashcode calculation and formation of products using, as an example, the formation of amides from acid chlorides and amines. The upper box shows how acid chlorides and amines in the reagent list are marked with X's, with the elimination of atoms that are not found in the product. The middle box shows the calculation of hashcodes for a sample acid-chloride and a sample acyclic amine. In calculating the hashcode for the acid-chloride, note that the external acyclic substituents, (methyl and chloro) are clipped off, whereas the markup atom X1 is retained. Also note that the double-bonded carbonyl oxygen connected directly to the bridge is saved. This reduced structure is then used to calculate the hashcode, which turns out to be DIZJYKQ. In the case of the amine, because of the absence of rings, all atoms except the X1 are clipped off, giving rise to the acyclic hashcode AAAAAAA. The lower box shows how the two reagents are merged via the marked-up bonds to give the final product.

same scaffold-component structure and reagents with hashcodes that do not have adequate representation are discarded. The combination of sets with the same hashcode from each list produces libaries of ~100 compounds, all with the same ring-scaffold. The combination of reagents to form the final product is illustrated in the lower part of Figure 1.

An example

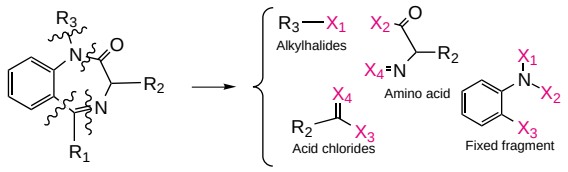
We illustrate the details of the design with an example from the literature – the synthesis of a benzodiazepine library [11]. The synthetic scheme involves a constant support-bound arylstannane, and three lists of reagents that introduce structural variability to the end-product, that is, alkyl halides (alkylating agents), Fmoc-protected amino acids and acid chlorides. Component-hashcodes for all the reagents were calculated, and reagent-sets with at least four or five instances of a shared hashcode were identified. The top part of Figure 2 shows the markup of these sets.

The results of the markup are shown in the lower part of Figure 2. These are grouped into sets of five (or fewer) reagents that share the same hashcode. For example, there are a total of eight reagents having the ring-scaffold hashcode LBCSVZQ. These are grouped into a set of five and a set of three. The hashcode EQXGNRK has only three representatives.

The top part of Figure 3 shows how the marked-up reagents are fused to form the final product. Sets of reagents from the three lists were combined to build 125-compound mini-libraries. For example, one might combine the set of five acid chlorides with the hashcode LBCSVZQ with five alkyl halides with the hashcode HNXHYYF and five Fmoc-protected amino-acids with the hashcode JVGWEID; this gives rise to 125 compounds with the same ring-scaffold but different acyclic substitutents. Combining other sets of reagents in the same manner could produce additional mini-libraries – with the same chemistry we produce several mini-libraries, each representing a different scaffold. The lower part of Figure 3 shows some of the products from a 125-member mini-library made via this procedure. It can be seen that all the structures share the same ring-scaffold and have different combinations of side-chain functional groups.

Discussion

Combinatorial chemistry is just one means of library expansion. The obvious benefit of an in-house combinatorial effort is the proprietary advantage. Unlike earlier work, where diversity is the central if not sole criterion for design, we have suggested making uniform libraries consisting of 100-compound sets where each set elaborates a single scaffold. To put this number in perspective, a different analysis by the Martin group [12] suggests that compounds within an 85% similarity cutoff of an active compound are only ~30% likely to be active themselves. This result cannot be compared directly with ours because our calculations are based on using the ring-scaffold as a basis for grouping compounds into series, while the Martin work uses a similarity calculation based on Unity fingerprints (UNITY is a product of Tripos,1699 South Hanley Road, St. Louis, Missouri 63144, USA). Nevertheless, their paper highlights the fact that although two compounds might be similar topologically, their biological activities could differ a lot. It is



Acid Chlorides

Ring-Scaffold	Scaffold With Branchpoints	MFCD
LBCSVZQ	*c1cccc(*)c1	00128835
LBCSVZQ	*c1ccc(*)cc1	00013257
LBCSVZQ	*c1cc(*)cc(*)c1	00128837
LBCSVZQ	*c1ccc(*)c(*)c1	00128836
LBCSVZQ	*c1ccccc1	00000637
LBCSVZQ	*c1ccc(*)cc1	00000638
LBCSVZQ	*c1ccc(*)cc1	00013255
LBCSVZQ	*c1ccc(*)cc1	00013258
NTTINZY	*c1cccnc1	00012820
NTTINZY	*c1cccnc1*	00051678
NTTINZY	*c1ccc(*)nc1	00051775
NTTINZY	*c1cnc(*)c(*)c1	00204230
NTTINZY	*c1cccnc1*	00051677

Alkylhalides

Ring-Scaffold	Scaffold With Branchpoints	MFCD
AAAAAAA	*Acyclic*	00064575
AAAAAAA	*Acyclic*	00000250
AAAAAAA	*Acyclic*	00190233
AAAAAAA	*Acyclic*	00000241
AAAAAAA	*Acyclic*	00009886
HNXHYYF	*c1cccc(*)c1*	00042503
HNXHYYF	*c1ccc(*)c(*)c1	00236025
HNXHYYF	*c1cc(*)cc(*)c1	00013539
HNXHYYF	*c1ccccc1*	00013558
HNXHYYF	*c1ccc(*)cc1	00085007
HNXHYYF	*c1cccc(*)c1	00000340
HNXHYYF	*c1ccc(*)cc1	00000179
HNXHYYF	*c1ccc(*)cc1	00000180

Fmoc-protected amino-acids

Ring-Scaffold	Scaffold With Branchpoints	MFCD
LGKORCD	*c1ccc(*)cc1	01317005
LGKORCD	*c1ccccc1	00155632
LGKORCD	*c1ccc(*)cc1	01320845
LGKORCD	*c1ccccc1	00190891
LGKORCD	*c1ccc(*)cc1	01320876
LRHKJGH	*Acyclic*	00235820
LRHKJGH	*Acyclic*	00235842
LRHKJGH	*Acyclic*	00235893
JVGWEID	*C1CCCCC1	01311779
JVGWEID	*C1CCCCC1	00190878
JVGWEID	*C1CCCCC1	01311752
JVGWEID	*C1CCCCC1	00190877
JVGWEID	*C1CCCCC1	00274784

Drug Discovery Today

Figure 2. The top part of the figure shows reagent markup. The connection points of the pieces (fixed fragment, alkyl halides, amino-acids and acid chlorides) are marked with X1, X2 and so on. These marked-up fragments are then used to calculate ring-scaffold hashcodes. The lower part of the figure shows a sample of the resulting hashcodes for the alkyl halides, amino-acids and acid chlorides. The 'ring-scaffold' column shows the 7-letter hashcode. The column titled 'scaffolds with branch points' contains a SMILES string [9] that indicates the scaffold with attachment points. The 'MFCD' column shows the MFCD number of the corresponding reagent (MFCD numbers are identification numbers in the Available Chemicals Directory database a product of MDL Information Systems Inc., 1460 Catalina Street, San Leandro, CA 94577, USA).

well known that a small change, such as the addition or deletion of a methyl group, can cause an active compound to become completely inactive [13].

It could be argued that scaffolds with a large number of open valences need more analogs to explore their scope than smaller scaffolds with few open valences. This is indeed a valid criticism in light of which the number of analogs actually made around each scaffold could be adjusted (upward or downward). An average value to be used as a general guide for the design of libraries is 100.

Another question might be whether it is better to synthesize, for example, 10,000 mutually dissimilar compounds, or 100 analogs each of 100 different scaffolds. Both would be 10,000 compound sets but the mutually dissimilar would cover more 'chemical space'. However, each series would not be adequately probed because it is represented by only one compound. The approach recommended here (100 analogs of 100 different scaffolds), is advantageous because it gives fair representation of each series and, additionally, might generate preliminary SAR after biological screening. The production of 100-analog sets requires robust synthetic methods but, once developed, can be re-used to rapidly synthesize analogs, should the series prove active at a future date.

Conclusions

Three main points are made here:

(1) A new type of screening library is proposed, which is 'uniform' in composition with 'equal' representation of different medicinally relevant scaffolds.

(2) Each scaffold has ~100 representatives (analogs). This number is derived from an analysis of historic HTS data.

(3) A practical method to assemble such combinatorial libraries has been developed and is effective.

The details of the implementation, for example, the use of the ring-scaffold as a surrogate for chemical series, are less important. The central take home message is the concept of a 'uniform library'. We hope that this fresh approach will attract the attention of researchers in both academia and industry and that it will be adopted as a beneficial option.

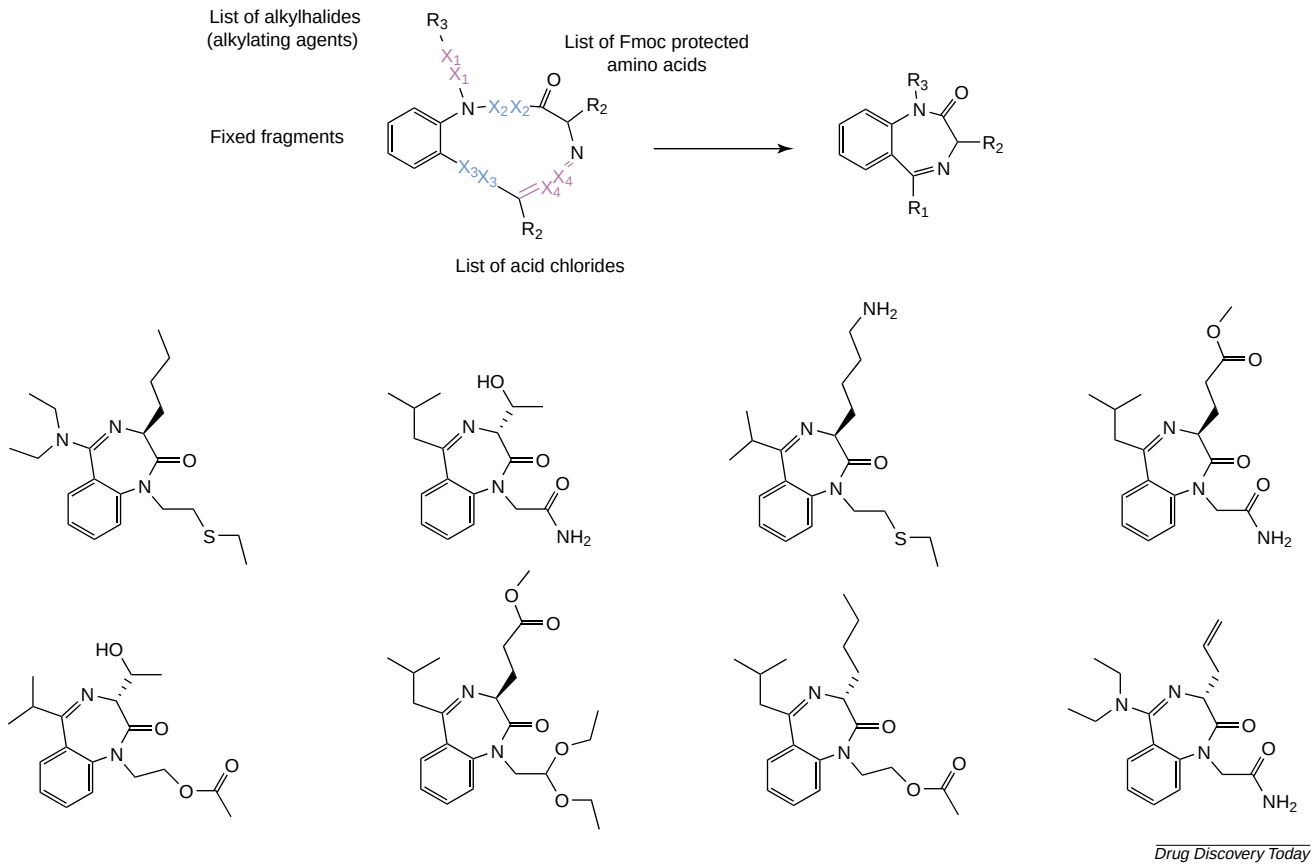


Figure 3. The upper part shows a schematic of how the marked-up reagents are fused to form the final product. Note, for example, that the two X1's disappear resulting in a single bond between the two atoms originally connected to the X1s. The lower part of the figure shows a sample of the resulting products.

Acknowledgements

Ramaswamy Nilakantan would like to thank Fred Immermann and Kevin Haraki for useful discussions, and John Ellingboe and Dominick Mobilio for enthusiastically supporting this project.

References

1 Good, A.C. and Lewis, R.A. (1997) A new methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* 40, 3926–3936

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

3 Wright, T. et al. (2003) Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 43, 381–390

4 Koehler, R.T. and Villar, H.R. (2000) Design of screening libraries biased for pharmaceutical discovery. *J. Comput. Chem.* 21, 1145–1152

5 Beno, B.R. and Mason, J.S. (2001) The design of combinatorial libraries using properties and 3D pharmacophore fingerprints. *Drug Discov. Today* 6, 251–258

6 Brown, R.D. et al. (2000) Combinatorial library design for diversity, cost efficiency and drug-like character. *J. Mol. Graph. Model.* 18, 427–437

7 Martin, E.J. and Critchlow, R.E. (1999) Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* 1, 32–45

8 Nilakantan, R. et al. (2002) A novel approach to combinatorial library design. *Comb. Chem. High Throughput Screen.* 5, 105–110

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

10 Nilakantan, R. et al. (1997) Database diversity assessment: New ideas, concepts, and tools. *J. Comput. Aided Mol. Des.* 11, 447–452

11 Bunin, B.A. et al. (1996) Synthesis and evaluation of 1,4-benzodiazepine libraries. In *Methods in Enzymology* (Vol. 267) (Abelson, J.N., ed.), pp. 448–465, Academic Press

12 Martin, Y.C. et al. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358

13 Schneider, G. (2000) Trends in virtual combinatorial library design. *Curr. Med. Chem.* 9, 2095–2101