CURRENT OPINION
www.current-opinion.com

Reagent-based and product-based computational approaches in library design

Eric A Jamois

The design of combinatorial libraries involves the consideration of all synthesizable compounds (the virtual library), followed by the selection of a suitably sized subset for actual synthesis and experimentation. Several approaches to this task can be envisaged, involving either reagent-based or product-based considerations. Reagent-based design considers the properties of the building blocks rather than those of the final products. Although popular with chemists, this approach overlooks the extent of chemical transformations involved in generating products. In effect, several important properties cannot be derived from building blocks alone and require access to product structures. Several studies have demonstrated the superiority of product-based designs in yielding diverse and representative subsets. Although more computationally intensive, the latter approach provides a basis for more sophisticated designs where reagent-based and product based considerations can be combined for a best-of-breed approach.

Addresses
Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA
e-mail: ericj@accelrys.com

Current Opinion in Chemical Biology 2003, **7**:326–330

This review comes from a themed issue on
Combinatorial chemistry
Edited by Samuel Gerritz and Andrew T Merritt

1367-5931/03/$ – see front matter

DOI 10.1016/S1367-5931(03)00053-X

Introduction

In the past few years, the efficient design of combinatorial libraries has become increasingly important in lead discovery and follow-up programs [1•,2,3•,4]. Through the advent of newly discovered reactions and new reagents available for purchase, the size of virtual libraries (total pool of synthesizable compounds) has increased dramatically over the past few years. (There are several sources of reagents for chemical synthesis, the most common one is MDL ACD. The database currently holds over 250 000 chemicals and grows at a rate of ca. 20 000 compounds per year.) Because of experimental constraints of synthesis and screening equipment, only a small fraction of this virtual pool can usually be tested. Consequently, these large libraries are usually reduced to smaller subsets. The challenge is to provide a selection of building blocks that best suits our needs in terms of product properties (e.g. diversity or focusing, drug-likeness, etc.) but also includes practical reagent-based considerations (reactivity, selectivity, cost, etc), possibly including other criteria, taken either individually or in combination.

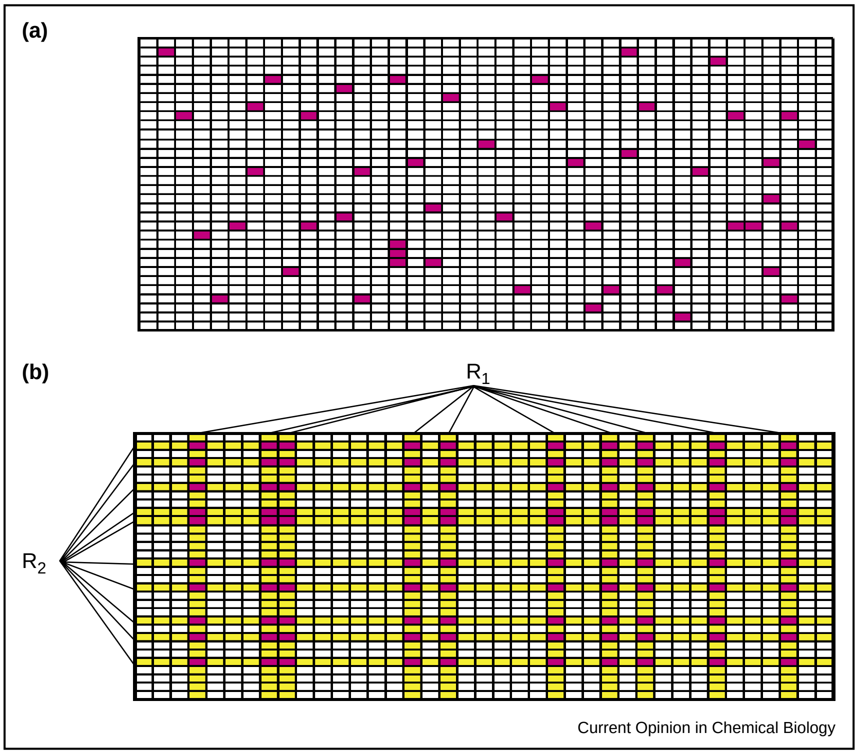
Several techniques have been used for the identification of library subsets [5–11]. One category of techniques involves reagent-based selections; that is, selections involving reagent properties only. Selection criteria may involve such diverse factors as reactivity, selectivity, cost, substructure and also 3D pharmacophore information [12,13•]. A second class of techniques involves the selection of diverse sets of products. Here, a range of dissimilarity-based [6–10] and clustering [11] methods has been used for selections at the product level. One of the limitations of dissimilarity-based methods applied to combinatorial libraries is the lack of combinatorial constraints that results in synthetically inefficient subsets (Figure 1). A third type of technique involves combinatorially constrained product selections. In this case, the combinatorial array is maintained by the selection of reagents but the evaluation of diversity of the resulting subset is performed at the product level. Such a procedure using genetic algorithms or Monte-Carlo optimization has been previously described [14,15,16••,17•,18••,19••]. Recently, techniques combining multiple optimization criteria such as diversity, cost efficiency and drug-like character have also appeared [16••,18••,19••]. It is therefore possible to combine the practical aspects of reagent selection or reagent bias with the more rigorous product-based approaches.

One major distinguishing feature between the techniques is the magnitude of the computational task at hand. Most reagent-based selection techniques require little computational resource because of the limited size of reagent lists. In fact, many selections have been performed through visual inspection of the reagent lists. Also, the size of the problem is only additive with respect to the size of each list. On the other hand, product-based techniques often require complex and time-consuming procedures because of the multiplicative nature of the problem. A reagent array of 50 × 150 × 200 × 350 for a four substituent system R1 × R2 × R3 × R4 would generate 525 million products. Both enumerative and non-enumerative solutions to this problem have been proposed.

Reagent-based design

Reagent-based design has been in practice among chemists for many years, its practical appeal and efficiency cannot be denied. In recent years, chemists have been selecting building blocks that incorporate drug-like fragments, with the hope that these would result in drug-like

Figure 1



Depiction of the benefits of combinatorial constraints. **(a)** A non-combinatorial selection that would be synthetically inefficient. **(b)** A selection with combinatorial constraints, which would therefore be much easier to implement via combinatorial synthesis.

molecules. Chemists have also learned to avoid building blocks lacking regio- or chemo-selectivity and yielding mixtures of products. Both of these procedures can be performed on the reagent lists using simple substructure considerations. Additional considerations involve cost and availability where reagents that are inexpensive and/or already in inventory should be selected preferentially. Several reagent-based design strategies have been reported and illustrate the popularity of this approach [12,13•,20].

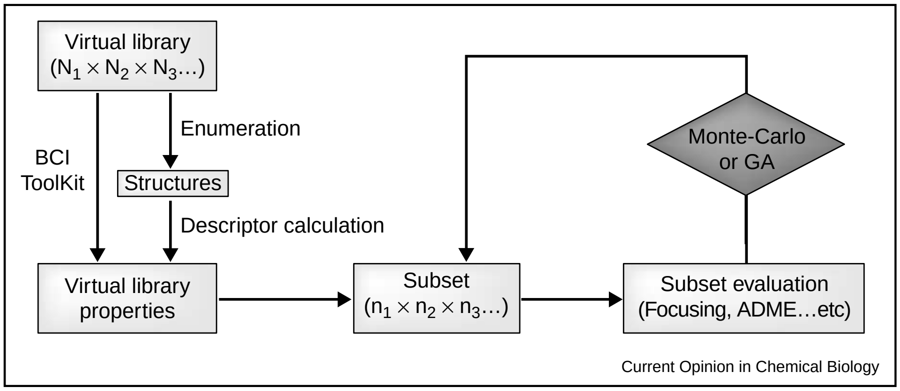
Reagent-based design can also be applied to generate diverse or, in some cases, focused subsets based on bioisosteric replacement [21]. Several methods have been used in the selection of monomers, including maximum dissimilarity [6], D-optimal design [21] and clustering [15]. Hierarchical cluster analysis has been well validated in compound selection [11], it has also provided acceptable results in our earlier work [15]. Although it does not provide an optimal solution, the method is fast and easy to implement: several sets of descriptors can be considered for analysis, ranging from MDL ISIS, Daylight or BCI fingerprints to physicochemical descriptors [22].

The limitations of reagent-based selections have been pointed out — considerations at the reagent level do not directly reflect the nature of the corresponding products [14,15,23•]. Although better than random, selections performed at the reagent level are sub-optimal in terms of diversity because of the non-additive nature of the descriptors involved. It has been shown that better selections can be obtained using product-based considerations [14,15]. The incorporation of drug-like properties follows a similar logic. Although it is possible to select building blocks with drug-like considerations, the extent of synthetic transformations makes this approach also sub-optimal. It is possible that a building block can be considered unsuitable when taken individually but that, in combination with other reagents, results in acceptable drug-like products. Each reagent should therefore be examined in its possible combinations with all other reagents and evaluated according to the drug-likeness of the corresponding products. Consequently, it is advisable that even simple drug-like rules such as the Lipinski rule of five [24] be applied at the product level. More complex models involving computational ADME do not usually lend themselves to reagent-based considerations and should realistically be applied at the product level [25,26•].

Product-based design

Product-based library design involves a more complex optimization procedure that we term 'combinatorial

Figure 2



Conventional library optimization workflow involving enumeration of the complete virtual library followed by optimization of a suitable subset. Early models for absorption, distribution, metabolism and excretion (ADME) can be used to bias the subset selection towards suitable compounds. A genetic algorithm (GA) can be used to optimize a population of libraries rather than a single individual. Barnard Chemical Information Ltd. (BCI) software provides Markush-based representation of combinatorial libraries for extremely fast descriptor calculation [32,33] (http://www.bci.gb.com).

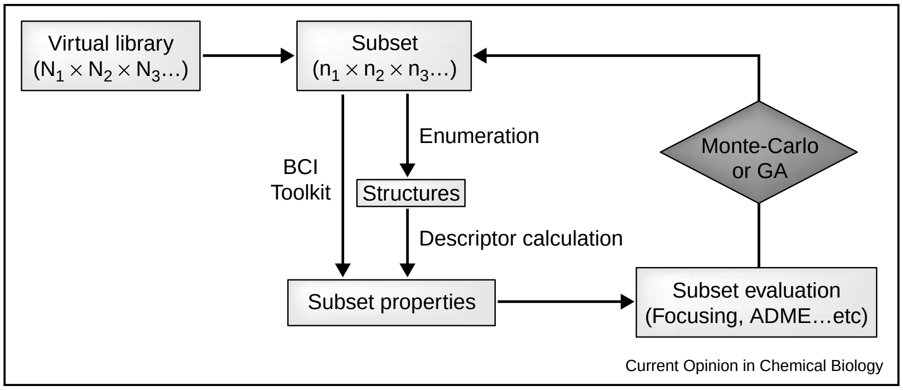
optimization' where the reagent selection is optimized against the properties of the corresponding products. In this scheme, the combinatorial nature of the sub-library is maintained through combinatorial constraints whereas evaluation of diversity, focusing or other criteria is performed on the products. Procedures using either a Monte Carlo optimization or genetic algorithms have been reported [14–16,17•,18••,19••]. These procedures are more computationally intensive than simpler reagent-based considerations because conventional techniques require full enumeration of the products, descriptor calculation for the entire library and optimization of the subset (Figure 2).

The combinatorial optimization process described above attempts to identify a selection of reagents, which provides the desired product properties. If we take the example of a virtual library consisting of a 100 × 100 × 100 array (1 000 000 possible products) for R1 × R2 × R3 and seek to isolate subsets of 10 × 10 × 10 (1000 compounds), the total number of possible solutions is $C_{100}{}^{10} \times C_{100}{}^{10} \times C_{100}{}^{10} = 5 \times 10^{39}$. This is a formidable number, which makes it generally impossible to systematically investigate every possible subset. So we rely on the optimization procedure to provide a near optimal solution. Related studies using genetic algorithms or the Monte-Carlo procedure described above suggest that the subsets obtained with such procedures are only slightly sub-optimal [14,15]. Several conditions have been reported for the Monte Carlo optimization using a large number of steps at a given temperature or a simulated annealing procedure [15,18••]. We have previously ascertained the consistent quality of the subsets returned, although the solutions themselves may be different [15].

Several molecular descriptors have been used in library design and in the characterization of molecular diversity or focusing [21,22]. Similar descriptors to those described earlier in the reagent-based design section are typically used in the analysis of products. Cell-based approaches typically provide fast measurements of product space coverage. However, a low dimensionality space is required. Several techniques are routinely used for dealing with the high dimensional problems. In the case of physicochemical descriptors, principal component analysis is routinely applied. In the case of fingerprint descriptors (MDL ISIS Keys, Daylight Fingerprints or BCI Fingerprints), diversity metrics such as those counting the number of on-bits can successfully be applied. For focused designs, the Tanimoto distance to a given lead compound can be used as a basis for optimization. Pharmacophore fingerprints have also been applied in library design scenarios [27,28].

Whereas early computational methods for library subsetting concentrated on either diversity or similarity to known leads, other aspects such as drug-likeness are also part of the library-design process. In this fashion, we can produce libraries whose hits can be more easily optimized into successful drug candidates. This approach requires that we pursue several objectives simultaneously; diversity (or similarity) and drug-likeness for example. It has been shown that drug-likeness can be achieved with minimal impact on the diversity of the compounds selected [18••]. So we can obtain sets that are almost equally diverse but different in their drug-like character. Such an outcome is not entirely surprising when we look at the vast number of possible sub-libraries that can be generated. The large ensemble of solutions, at first perceived as a liability, may be turned into an asset where we

Figure 3



On-the-fly library optimization workflow.

now have the flexibility to provide libraries that are both drug-like and diverse.

Ideally, we would like to apply product-based design criteria for properties that are not adequately represented by the reagents (such as diversity and drug-like character) and reagent-based design criteria for practical considerations such availability, cost or ease of handling. For example, in addition to diversity and drug-like character, we would like to introduce a bias so that reagent selection is directed towards those preferred by chemists. Such a process involves simultaneous optimization against several criteria that need to be balanced appropriately. If the criteria involved in the optimization are normalized, then we can ensure the proper balance between them and reach an acceptable compromise solution. In this fashion, we can provide sub-libraries that are combinatorial, reasonably diverse, drug-like and use mostly 'desirable' reagents. A procedure involving such conditions has been reported recently [18••].

As mentioned earlier, product-based library design introduces considerable computational complexity, involving library enumeration, descriptor calculation and the optimization procedure itself. The latter can be performed rather efficiently so only the former two steps remain as bottlenecks. Novel approaches have appeared in the literature involving partial enumeration [29•], stochastic sampling [30•] and GA-based optimization of subsets [31•]. For descriptor calculation, Markush-based representations have been used to handle a number of 2D descriptors with order of magnitude speedup compared to conventional approaches [32,33]. We recently experimented with an 'on the fly optimization' workflow (Figure 3). This modified workflow combines the optimization of subsets using Monte-Carlo or GA with a Markush-based descriptor calculation [34]. Results suggest that suitable solutions may be obtained while sampling only a very small fraction of the complete virtual library, and confirm earlier findings [29•–31•]. These breakthroughs make product-based library design more efficient, especially when dealing with very large virtual libraries.

Conclusion

One of the challenges in modern library design is to involve both theoretically sound methods and practical considerations. Theoretical bases rely primarily on computational chemists whereas practical aspects are dictated by experimental chemists. The implementation of a successful library design strategy draws expertise from these two groups and conciliates the appeal of theory with the reality of experiments. Computational design as we know it has seldom produced the final answer to a library design problem, as 'touch ups' are often required. For example, it is common to provide replacements for reagents that were initially selected but, for some reason (commercial availability or other), cannot be obtained or used. Thanks to computational methods, we can make educated suggestions for such replacements. We can also evaluate alternative propositions, such as those provided by experimental chemists, and ensure that the resulting library retains most or all the characteristics that were originally intended.

Acknowledgements

I acknowledge Dr Marvin Waldman and his group at Accelrys for helpful scientific discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. • Ghose AK, Viswanadhan VN (eds): *Combinatorial Library Design and Evaluation – Principles, Software Tools and Applications in Drug Discovery*. New York: Marcel Dekker; 2001.
Provides a recent and comprehensive coverage of the topic with many references.

2. Valler MJ, Green D: **Diversity screening versus focused screening in drug discovery**. *Drug Discov Today* 2000, **5**:286-293.

3. Pickett SD, McLay IM, Clark DE: **Enhancing the hit to lead**
• **properties of lead optimization libraries**. *J Chem Inf Comput Sci* 2000, **40**:263-272.

Provides an interesting example of incorporating 'drug-likeness' in the library design process.

4. Martin EJ, Crichlow RW: **Beyond mere diversity: tailoring combinatorial libraries for drug discovery**. *J Comb Chem* 1999, **1**:32-45.

5. Johnson M, Maggiora GM: *Concepts and Applications of Molecular Similarity*. New York, NY: Wiley; 1990.

6. Snarey M, Terrett NK, Willett P, Wilton DJ: **Comparison of algorithms for dissimilarity-based compound selection**. *J Mol Graphics Mod* 1997, **15**:372-385.

7. Lajiness MS: **Dissimilarity-based compound selection techniques**. *Perspect Drug Discovery Design* 1997, **7**:65-84.

8. Marengo E, Todeschini R: **A new algorithm for optimal distance-based experimental design**. *Chemo-metrics Intell Lab Syst* 1992, **16**:37-44.

9. Holliday JD, Ranade SS, Willett P: **A fast algorithm for selecting sets of dissimilar structures from large chemical databases**. *Quant Struct-Activity Relationships* 1995, **14**:501-506.

10. Hassan M, Bielawski JP, Hempel JC, Waldman M: **Optimization and visualization of molecular diversity of combinatorial libraries**. *Mol Divers* 1996, **2**:64-74.

11. Brown RD, Martin YC: **Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection**. *J Chem Inf Comput Sci* 1996, **36**:572-584.

12. Zheng W, Cho SJ, Tropsha A: **Rational combinatorial library design. 1. Focus-2D: a new approach to targeted combinatorial chemical libraries**. *J Chem Inf Comput Sci* 1998, **38**:251-258.

13. Leach AR, Green DVS, Hann MM, Judd DB, Good AC: **Where are**
• **the gaps? A rational approach to monomer acquisition and selection**. *J Chem Inf Comput Sci* 2000, **40**:1262-1269.

Provides an interesting example of reagent-based design with 3D pharmacophore descriptors.

14. Gillet VJ, Willett P, Bradshaw J: **The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries**. *J Chem Inf Comput Sci* 1997, **37**:731-740.

15. Jamois EJ, Hassan M, Waldman M: **Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets**. *J Chem Inf Comput Sci* 2000, **40**:63-70.

16. Gillet VJ, Willett P, Bradshaw J, Green DVS: **Selecting**
•• **combinatorial libraries to optimize diversity and physical properties**. *J Chem Inf Comput Sci* 1999, **39**:169-177.

A reference of special interest that details how optimization can be performed against multiple objectives.

17. Reynolds CH, Tropsha A, Pfahler LB, Druker R, Chakravorty S,
• Ethiraj G, Zheng W: **Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms**. *J Chem Inf Comput Sci* 2001, **41**:1470-1477.

Comparison of several compound selection algorithms and schemes.

18. Brown RD, Hassan M, Waldman M: **Combinatorial library design**
•• **for diversity, cost efficiency, and drug-like character**. *J Mol Graphics Mod* 2000, **18**:427-437.

A reference of special interest with optimization against multiple objectives. More specific to the Cerius2 software suite.

19. Gillet VJ, Khatib W, Willett P, Fleming PJ, Green DVS:
•• **Combinatorial library design using a multiobjective genetic algorithm**. *J Chem Inf Comput Sci* 2002, **42**:375-385.

A reference of special interest that details how one may deal with potentially conflicting design objectives and how to explore choices for a compromise solution.

20. Lewell XQ, Judd DB, Watson SP, Hann MM: **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* 1998, **38**:511-522.

21. Langer T, Hoffmann RD: **New principal components derived parameters describing molecular diversity of heteroaromatic residues**. *Quant Struct-Act Relat* 1998, **17**:211-223.

22. Brown RD: **Descriptors for diversity analysis**. *Perspect Drug Discov Des* 1997, **7**:31-49.

23. Leach AR, Bradshaw J, Green DVS, Hann MM: **Implementation of**
• **a system for reagent selection and library enumeration, profiling, and design**. *J Chem Inf Comput Sci* 1999, **39**:1161-1172.

Illustrates how library design tools can be integrated with other computational and cheminformatics components and deployed to chemists.

24. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Adv Drug Delivery Rev* 1997, **23**:3-25.

25. Eagan WJ, Merz KM, Baldwin JJ: **Prediction of drug absorption using multivariate statistics**. *J Med Chem* 2000, **43**:3867-3877.

26. Darvas F, Dorman G, Papp A: **Diversity measures for enhancing**
• **ADME admissibility of combinatorial libraries**. *J Chem Inf Comput Sci* 2000, **40**:314-322.

Illustrates how early ADME considerations in library design can improve success rate from *in vitro* hits to *in vivo* leads.

27. McGregor MJ, Muskal SM: **Pharmacophore fingerprinting. 1. Application to QSAR and focused library design**. *J Chem Inf Comput Sci* 1999, **39**:569-574.

28. McGregor MJ, Muskal SM: **Pharmacophore fingerprinting. 2. Application to primary library design**. *J Chem Inf Comput Sci* 2000, **40**:117-125.

29. Agrafiotis DK, Lobanov VS: **Ultrafast algorithm for designing**
• **focused combinatorial arrays**. *J Chem Inf Comput Sci* 2000, **40**:1030-1038.

Interesting example of fast design of focused libraries using partial enumeration.

30. Lobanov VS, Agrafiotis DK: **Stochastic similarity selections from**
• **large combinatorial libraries**. *J Chem Inf Comput Sci* 2000, **40**:460-470.

Provides some interesting conclusions on the quality of results that can be obtained with minimal sampling against extremely large virtual libraries.

31. Sheridan RP, SanFeliciano SG, Kearsley SK: **Designing targeted**
• **libraries with genetic algorithms**. *J Comput-Aided Mol Design* 2000, **18**:320-334.

Provides an elegant approach to the optimization of subsets from extremely large virtual libraries using a genetic algorithm with 2D or 3D objective functions.

32. Downs GM, Barnard JM: **Techniques for generating descriptive fingerprints in combinatorial libraries**. *J Chem Inf Comput Sci* 1997, **37**:59-61.

33. Barnard JM, Downs GM, Scholley-Pfab A, Brown RD: **Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries**. *J Mol Graphics Mod* 2000, **18**:452-463.

34. Cerius^2, Version 4.8, Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121, USA On World Wide Web URL: http://www.accelrys.com/cerius2/c2libx.html