Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2003) 141–149

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Design of focused and restrained subsets from extremely large virtual libraries

Eric A. Jamois*, Chien T. Lin, Marvin Waldman

Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA

Received 27 February 2003; received in revised form 29 May 2003; accepted 6 June 2003

Abstract

With the current and ever-growing offering of reagents along with the vast palette of organic reactions, virtual libraries accessible to combinatorial chemists can reach sizes of billions of compounds or more. Extracting practical size subsets for experimentation has remained an essential step in the design of combinatorial libraries. A typical approach to computational library design involves enumeration of structures and properties for the entire virtual library, which may be unpractical for such large libraries. This study describes a new approach termed as on the fly optimization (OTFO) where descriptors are computed as needed within the subset optimization cycle and without intermediate enumeration of structures. Results reported herein highlight the advantages of coupling an ultra-fast descriptor calculation engine to subset optimization capabilities. We also show that enumeration of properties for the entire virtual library may not only be unpractical but also wasteful. Successful design of focused and restrained subsets can be achieved while sampling only a small fraction of the virtual library. We also investigate the stability of the method and compare results obtained from simulated annealing (SA) and genetic algorithms (GA).
© 2003 Elsevier Inc. All rights reserved.

Keywords: Virtual libraries; Optimization cycle; Simulated annealing

1. Introduction

In the last few years, the efficient design of combinatorial libraries has become increasingly important in lead discovery and follow-up programs [1–4]. Through the advent of new reactions discovered and new reagents available for purchase, the size of virtual libraries (complete pools of synthetically accessible compounds) has increased dramatically over the last few years. Interestingly, the typical size of synthesized libraries has actually decreased over the same period. This trend acknowledges the desire to make focused rather than general purpose screening libraries. It also accounts for the more precise guidelines used in the design. In this process, practical size subsets are extracted from these large virtual libraries. The actual size of the subsets is usually defined according to available synthesis and screening resources. In most cases, synthesized subsets represent a very small fraction of the virtual library. The problem therefore becomes the selection of the reagents that provide a set of products consistent with the design objectives (diversity, focusing, drug-likeness . . . , etc.). In this endeavor, computational approaches provide a powerful handle on the design of combinatorial library subsets.

A number of techniques have been used towards the identification of library subsets. A first category of techniques involves reagent-based selections; that is, selections involving reagent properties only. Although popular with chemists, this approach masks the extent of chemical transformations involved in generating products. In effect, a number of important properties cannot readily be derived from building blocks alone and require access to product structures. A second type of technique involves combinatorially constrained product selections. In this case, the combinatorial array is maintained through selection of reagents but the evaluation of the resulting subset properties is performed at the product level. Such a procedure using Monte Carlo (MC) or genetic algorithms (GA) has been previously described [5–10]. Several studies have demonstrated the superiority of product-based design [5,6]. Although more computationally intensive, the latter approach provides a basis for more sophisticated designs where reagent- and product-based considerations can be combined for a best of breed approach.

One major distinguishing feature between the techniques is the computational resource required to identify library subsets. Most reagent-based selection techniques require

* Corresponding author. Tel.: +1-858-799-5514; fax: +1-858-799-5100.
E-mail address: ericj@accelrys.com (E.A. Jamois).

1093-3263/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1093-3263(03)00154-2

little computational resource due to the limited size of reagent lists. Also, the size of the problem is only additive with respect to the size of each reagent list. On the other hand, product-based techniques often require complex and time consuming procedures due to 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. While the reagent lists can be handled separately by conventional analysis techniques, handling the full set of products through such techniques poses a serious challenge. This complexity has sometimes been circumvented through partial enumeration, stochastic, or coverage-based sampling techniques [11–15]. We offer an approach similar to the one described by Sheridan and co-workers [13,14] wherein subsets are extracted and descriptors calculated as needed to perform the desired optimization. However, our approach takes advantage of high-throughput ADME models as well as extremely fast descriptor calculation available via the BCI ToolKit which does not require intermediate enumeration of structures.

2. Background

2.1. Workflow in library design

The conventional workflow in library design involves enumeration of structures and calculation of molecular descriptors for the entire virtual library (Fig. 1a). In the case of extremely large virtual libraries, the requirements for this task may exceed available computational and storage resources. For example, we have estimated that storing only the four Lipinski properties [16] for a library of 537 million compounds would require around 30 GB of disk space [17]. Larger libraries or more complex sets of descriptors would require more. We would like to bypass these limitations and decouple the complexity of the library optimization task from the size of the virtual library. This may be accomplished through a different workflow (Fig. 1b) where descriptors are computed as needed within the subset optimization cycle and, in some cases, without intermediate enumeration of structures.

2.2. Descriptor calculation

A number of descriptors have been used in library design [18–21] ranging from MDL ISIS, Daylight or BCI fingerprints to topological indices. Significant progress has been made in the calculation of descriptors for library design. For example, via calculation of *A* log *P*98 and fast approximation of polar surface area (PSA), new avenues have been opened for high-throughput ADME models which can be used early in the design process [17,22]. Other advances include the ability to compute properties such as two-dimensional (2D) fingerprints, structural and topological descriptors directly from Markush structures [23,24]. In this case, properties are computed directly from the encoded library (RG file) without enumerating structures. The latter approach provides several orders of magnitude speedup compared to a conventional scheme and is particularly well suited for the previously described workflow (Fig. 1b). It is best applied to additive or semi-additive descriptors such as MW, H bond acceptor, H bond donor, log *P*, 2D fingerprints and topological descriptors (Kier and Hall connectivity and shape indices, subgraph counts and Zagreb index) where the most significant performance gains are achieved. Several other examples involving the use of decomposable descriptors in library design have appeared in the literature [25–27]. Neural networks have also been used in an attempt to relate building block properties to those of the resulting products [28].

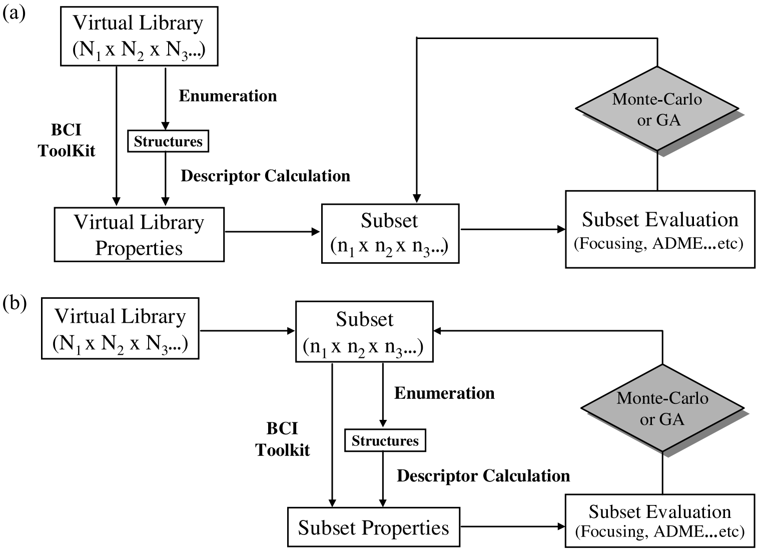
(a)
Virtual Library (N_1 x N_2 x N_3...)
Enumeration
BCI ToolKit
Structures
Descriptor Calculation
Virtual Library Properties
Subset (n_1 x n_2 x n_3...)
Subset Evaluation (Focusing, ADME...etc)
Monte-Carlo or GA
(b)
Virtual Library (N_1 x N_2 x N_3...)
Subset (n_1 x n_2 x n_3...)
Enumeration
BCI Toolkit
Structures
Descriptor Calculation
Subset Properties
Subset Evaluation (Focusing, ADME...etc)
Monte-Carlo or GA

Fig. 1. (a) Conventional library optimization workflow, (b) library optimization workflow with OTFO.

2.3. *Definition of property space*

The diversity or focusing characteristics of a sub-library can be evaluated using a number of distance-based or cell-based methods [29–32]. These methods, implemented as diversity metrics, evaluate how much of the complete library space is occupied by the subset or how well the subset focuses on a given lead compound. In a conventional workflow, the complete library space can be explicitly defined since the properties are enumerated for all compounds. This is unlikely to be the case in a situation where only a small fraction of the virtual library is sampled. Consequently, the workflow described (Fig. 1b) often relies on an approximation of the space in which the optimization is performed. However, a special situation arises when working with fingerprints such as those provided by MDL ISIS, Daylight or BCI. In this case, the descriptor is naturally bound since each bit in the fingerprint is set to either 0 or 1. Therefore, the definition of property space is inferred from the descriptor and does not rely on sampling of the virtual library.

2.4. *Subset evaluation*

The evaluation of derived subsets is required as an objective measure of their quality. The quality of a subset is defined as its ability to conform to the design objectives. For simple restrained designs, it refers to the ability of the subset to conform to the desired property distributions. For focused designs, we measure the ability of the subset to provide compounds which are similar to the target lead compound. Finally, for focused designs under restraints, we measure the ability of the subset to meet both of the previous objectives.

3. Methods

3.1. *Test libraries*

Two different combinatorial libraries were used in this work. The first library is based on the Tripeptoid library (Fig. 2a) described by Zuckermann et al. [33]. The building blocks are from 592 primary amines which are available from the available chemicals directory (ACD). The complete virtual library is of format 592 × 592 × 592 and contains 207 million compounds. The second library is an UGI library (Fig. 2b), similar to that described by Lobanov and Agrafiotis [12]. The building blocks are from 1442 acids, 592 primary amines, 37 aldehydes and 17 isonitriles also available from ACD. The complete virtual library is of format 1442 × 592 × 37 × 17 and contains 537 million compounds.

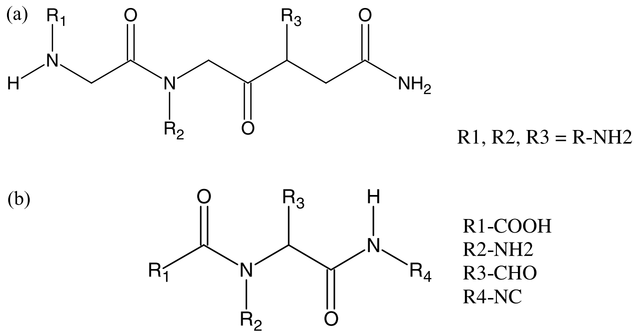

Fig. 2. (a) Tripeptoid library, (b) UGI library.

3.2. *Descriptors*

The first part of the study (Tripeptoid library) involves the absorption level calculated from $A\log P98$ and a fast approximation of PSA [17,22]. The second part of the study (UGI library) involves two sets of descriptors; BCI fingerprints (1052 bits) and structural Lipinski like descriptors (MW, H bond acceptor, H bond donor and $S\log P$) also available within the BCI ToolKit. The BCI fingerprints are used for focusing around a given lead compound. The structural descriptors are used to set restraints consistent with the guidelines provided by Lipinski et al. [16].

3.3. *Property restraints*

The first part of the study (Tripeptoid library) involves a simple design with restraints only. The predicted absorption level is used as a basis to define restraints (Table 1). The absorption levels are defined in four discrete intervals: 0 = good, 1 = moderate, 2 = poor, 3 = very poor. In this design, we attempt to select molecules which are predicted to be well absorbed while maintaining the combinatorial constraints. In the second part of the study (UGI library) property restraints involve structural Lipinski like descriptors (MW, H bond acceptor, H bond donor and $S\log P$). The restraints are used either alone or in combination with another objective function that attempts to focus on the lead compound. For each property, we define the minimum and maximum bounds interval inside which there is no penalty. We also define the standard deviation and the N-fold times standard deviation above the maximum bound at which the penalty reaches its capped maximum value. The handling of restraints in the Cerius^2 combinatorial chemistry software suite has been previously described [9,17].

Table 1
Conditions for restrained designs

Library/property Tripeptoid	Minimum	Maximum	S.D.	Maximum penalty N (Max + (N × S.D.))
Absoption level	0	0	1	3
UGI				
MW	0	500	100	4
H bond acceptor	0	10	2	4
H bond donor	0	5	1	4
$S\log P$	0	5	2	4

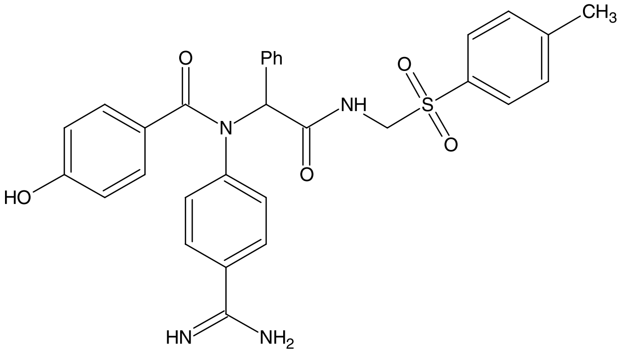

Fig. 3. Target lead (1.4 μM Thrombin inhibitor).

3.4. Focused design

The second part of the study (UGI library) involves focusing on a known Thrombin inhibitor (Fig. 3) [12]. In this case, the function optimized represents the Tanimoto distance between the library subset and the lead compound. The objective function is actually a composite of minimum and mean distances to the lead with equal weighting. As the virtual library contains the actual lead compound, we can evaluate if the resulting subsets also contain the lead.

3.5. Optimization protocols

3.5.1. Simulated annealing

The initial starting points consist of 10 random selections of the proper combinatorial format. These random selections are then optimized in an annealing procedure (Table 2). The temperature directly relates to the likelihood of a bad move (in this case, one that increases the objective function) being accepted. This likelihood decreases as the optimization progresses between $T = 1000$ and $T = 10$. Optimization is performed via mutations where one or two reagents (mutations column in Table 2) are exchanged for another one. A number of idle iterations (iterations with no progress made) are performed to ensure convergence of the optimization. With this setup, optimization will proceed with a minimum of 2500 iterations and a maximum of 10,000 iterations. The results are examined for consistency across the set of solutions to ensure reliability of the method.

3.5.2. Genetic algorithm

The initial population consists of 50 random individuals of the proper combinatorial format. This population is then optimized with the GA. We applied 20% crossover and 80% mutation with a maximum of 200 generations and 20 idle generations. The parent selection is performed by selecting the most fit individual from a random selection of 5.

Table 2
Simulated annealing procedure

Steps	Temperature	Iterations	Idle iterations	Mutations
1	1000	2000	500	2
2	300	2000	500	2
3	100	2000	500	1
4	30	2000	500	1
5	10	2000	500	1

3.6. Initial selections

The initial selection formats are 8 × 8 × 8 for the Tripeptoid library (512 compounds) and 12 × 6 × 4 × 3 for the UGI library (864 compounds). These formats are maintained in the course of the optimizations with simulated annealing (SA) or GA.

The initial random selections were analyzed to provide a baseline in our analysis. They were obtained by using the same seed as in the actual runs but with no optimization cycles performed. In this fashion, we could reproduce the original starting conditions in the optimizations and assess the stability of the results obtained from this methodology.

3.7. Subset evaluation

Evaluation of the identified library subsets is performed in several ways.

Objective functions: The objective functions used in the optimization provide interesting insight into the course of the optimization. We can evaluate if the optimization has converged and easily check on the consistency of the results. However, further analysis of the subsets is required to provide full insight into property distributions and other characteristics.

Property distribution: In the case of the Tripeptoid library, we can easily check that optimization of the subset indeed translates into suitable distributions of compounds across absorption levels and confidence ellipses of the Egan model [22]. In the case of the UGI library, we can plot distributions for the optimized properties and compare them to the random initial selection. We again ensure that optimization of the subset indeed translates into suitable distributions for the optimized properties.

Distance to lead compound: For the focused designs performed on the UGI library (with and without restraints), we can measure the Tanimoto distance to the target lead for each compound in an optimized subset. In this process, we produce a histogram of distances that can be compared to the same distances obtained on the random initial selection. We can also obtain D_{mean}, the mean distance to the target lead compound. This method has been used previously to control the quality of subsets [6].

4. Results

4.1. Tripeptoid library

A summary of the results for the Tripeptoid library is provided in Table 3 with objective function values

Table 3
Values of objective function for subsets optimized with SA and GA

Method	Initial libraries			Optimized libraries			
	Min.	Max.	Ave.	Min.	Max.	Ave.	Run time[a]
Restrained design							
SA[b]	2.951	5.928	4.680	0.000	0.000	0.000	5 h 30 min
GA[c]	1.891	7.000	4.479	0.000	0.000	0.000	1 h 00 min

[a] Silicon graphics, octane R10,000 195 MHz.
[b] Results over 10 individuals. Run time is for each individual.
[c] Results over 50 individuals for initial population and 45 individuals in final population.

Table 4
Values of objective function for subsets optimized with SA and GA

Method	Initial libraries			Optimized libraries			
	Min.	Max.	Ave.	Min.	Max.	Ave.	Run time[a]
Restrained design							
SA[b]	0.173	0.644	0.463	0.000	0.000	0.000	2 h 22 min
GA[c]	0.096	1.494	0.574	0.000	0.000	0.000	20 min
Focused design							
SA[b]	0.426	0.538	0.480	0.142	0.153	0.146	5 h 20 min
GA[c]	0.405	0.552	0.486	0.151	0.151	0.151	3 h 30 min
Focused design with restraints							
SA[b]	0.659	1.131	0.943	0.162	0.178	0.170	5 h 10 min
GA[c]	0.561	1.996	1.060	0.175	0.175	0.175	4 h 30 min

[a] Silicon graphics, octane R10,000 195 MHz.
[b] Results over 10 individuals. Run time is for each individual.
[c] Results over 50 individuals for initial population.

corresponding to the penalties for each subset. Only a restrained design was applied to this first library.

We now compare distributions of compounds across absorption levels and confidence ellipses for an initial random selection (Fig. 4a) and an optimized subset (Fig. 4b). We select the individual that is most representative of the quality of initial selections (objective function value closest to the set average). In this example, the individual has an initial objective function value of 4.623 and an optimized value of 0.000 as all other individuals in the optimized set of libraries.

4.2. UGI library

A summary of the results for the UGI library is provided in Table 4 with objective function values corresponding to (i) penalties (restrained design), (ii) distance function to the lead compound (focused design) and (iii) a composite function with both penalties and distance information to the lead compound.

For the best focused library obtained (objective function = 0.142), we investigate the distribution of Tanimoto distance to the lead compound. The histogram obtained (Fig. 5) reveals value of D_{mean} of 0.561 and 0.285 for the starting point in the optimization (blue) and the optimized subset (red), respectively.

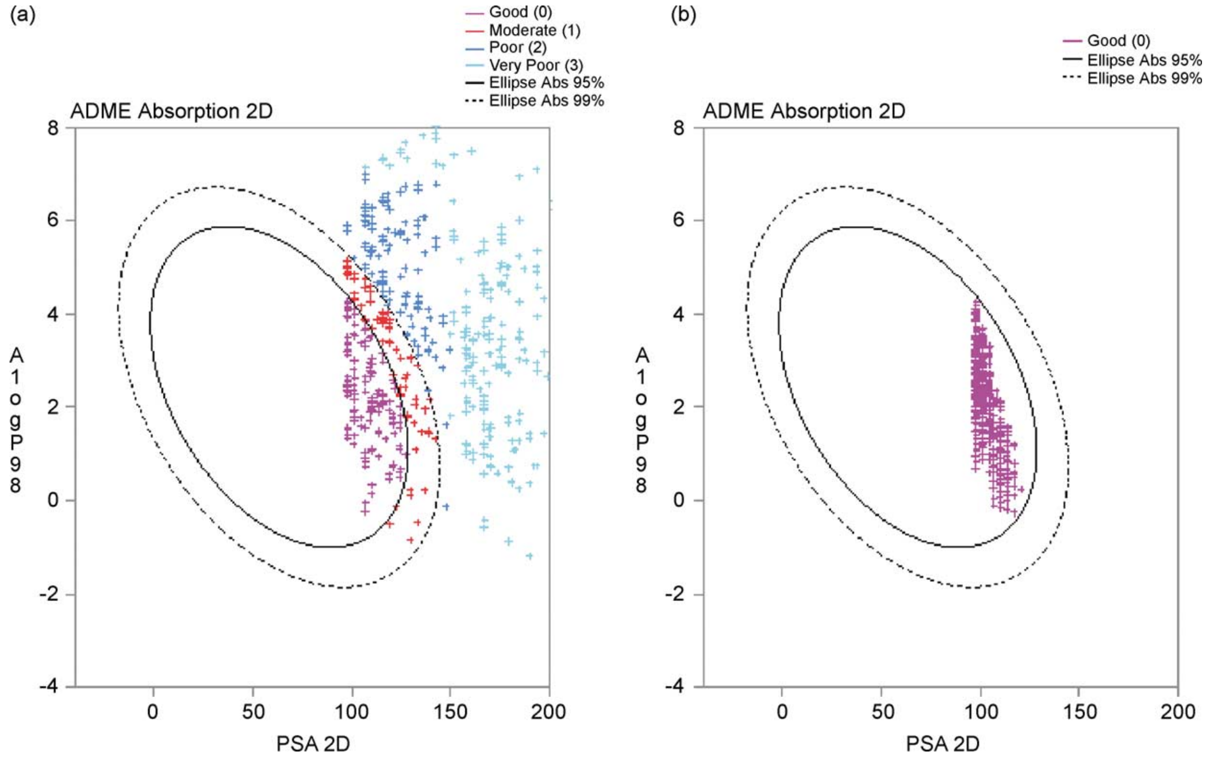

Fig. 4. Absorption plots for initial random selection (a) and optimized subset (b).

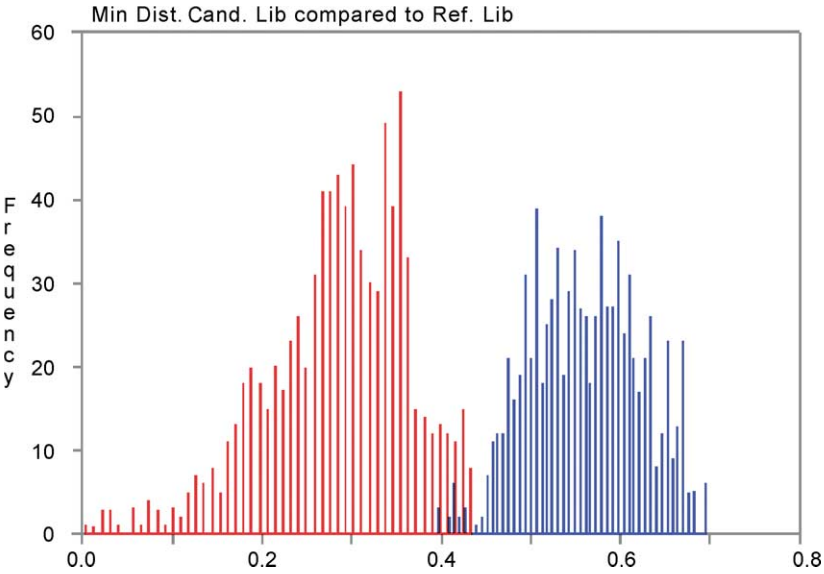

Fig. 5. Distance histograms to lead compound in focused design.

For this subset, we also analyze the molecular weight distribution. We compare the distribution obtained from the focused subset to that of the random and restrained subsets (Fig. 6). We observe that the focused subset contains a large proportion of large molecular weight compounds, although not as large as in the starting random subset. The restrained subset displays an ideal molecular weight distribution.

In an attempt to find a compromise between the different designs, we report a summary with the best results from the different optimization objectives (Table 5). We also cross-examine the subsets for properties that were not part of the optimization. For example, we measure the penalties against restraints for the focused library. We also measure the distance function to the lead compound for the restrained library.

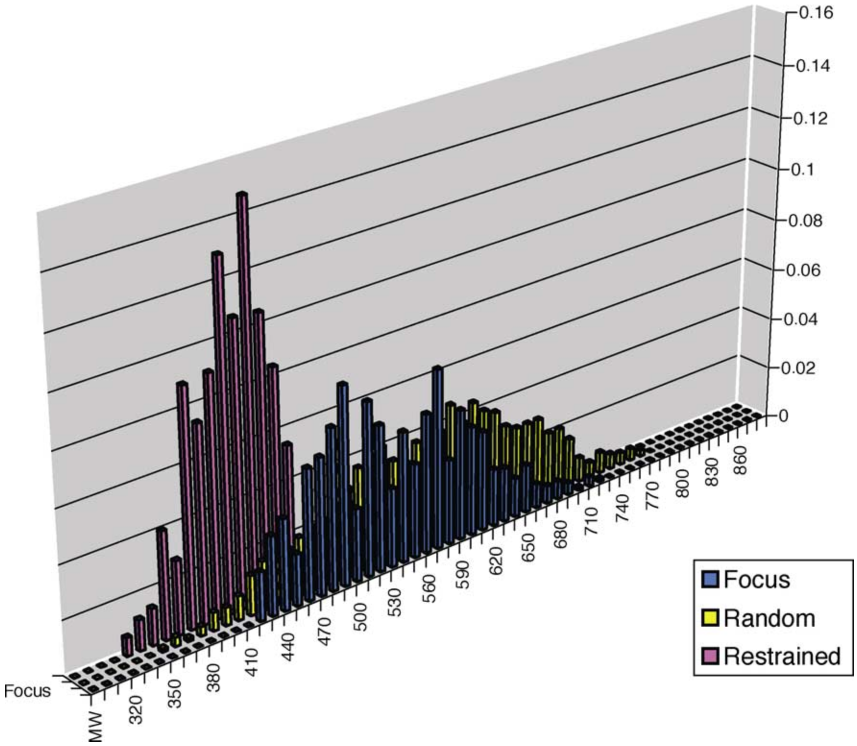

Fig. 6. Molecular weight distribution for random, focused and restrained subsets.

Table 5
Properties of optimized subsets for multiple design objectives

Design	Distance function to lead[a,b]	Penalty vs restraints[b]
Random	0.477	0.529
Focused design	**0.142**	0.198
Restrained design	0.483	**0.000**
Focused design with restraints	**0.152**	**0.010**

[a] Distance function to lead = 0.5 × minimal distance to lead + 0.5 × mean distance to lead.

[b] Optimized parameters are highlighted in bold face.

5. Discussion

5.1. Combinatorial optimization

The combinatorial optimization process attempts to identify a selection of reagents, which provides one or several libraries with the desired set of properties. In our test cases, the process optimizes an $8 \times 8 \times 8$ array from the complete array of $592 \times 592 \times 592$ (Tripeptoid library) or $12 \times 6 \times 4 \times 3$ from $1442 \times 592 \times 37 \times 17$ (UGI library). For the Tripeptoid library, the total number of possible $8 \times 8 \times 8$ subsets is $C_{592}^{8} \times C_{592}^{8} \times C_{592}^{8} = 4.5 \times 10^{52}$, for the UGI library, the total number of possible $12 \times 6 \times 4 \times 3$ subsets is $C_{1442}^{12} \times C_{592}^{6} \times C_{37}^{4} \times C_{17}^{3} = 4.2 \times 10^{50}$, both formidable numbers. Since, it would be impossible to systematically investigate every possible subset, we rely on the SA and GA procedures to provide near optimal solutions. Related studies using GAs suggest that the subsets obtained with such procedures are only slightly sub-optimal [5].

Given the complexity and size of our solution space, we can imagine a relatively flat fitness landscape (especially for simple design objectives), where a large number of equally good solutions may be found. In the vast ensemble of reagent combinations, we may be able to find selections that provide a large number of characteristics simultaneously. For example, we may be able to identify subsets that not only focus on a given lead compound but also present suitable distributions for Lipinski properties. Other design guidelines could be added, involving either product or reagent level considerations.

5.2. Optimization of Tripeptoid library

Optimization of the Tripeptoid library pursues a single objective: to select building blocks that lead to compounds falling within the most suitable absorption level (level 0). We observe that all random selections in the initial SA or GA population contain large percentages of compounds in the poor (level 2) and very poor (level 3) absorption levels. Optimization of starting selections to suitable subsets can be performed with SA or GA in a reliable fashion. In all cases, subsets are optimized to zero penalty with no dependence on the quality of the starting selection.

5.3. Optimization of UGI library

5.3.1. Restrained design

The restrained design attempts to select building blocks that lead to compounds that conform to the property guidelines (Table 1). We observe that all random selections in the initial SA or GA population contain some amount of violations against these guidelines. Optimization of starting selections to suitable subsets can be performed with SA or GA in a reliable fashion. In all cases, subsets are optimized to zero penalty with no dependence on the quality of the starting selection.

5.3.2. Focused design

The focused design attempts to select building blocks that lead to compounds similar to the target Thrombin inhibitor (Fig. 3). We observe that random selections have no focusing capabilities towards the lead compound. Optimization of starting selections to focused subsets can be performed with SA or GA in a reliable fashion. Results from SA provided 10 libraries of which 9 contained the actual lead compound. For the library that did not contain the actual lead, three out of the four R group positions (R2, R3 and R4) correctly mapped to the lead compound building blocks. The latter library ranked ninth out of the 10 results from SA; it is therefore unlikely that it would be selected as the preferred solution. Results from GA provided two libraries which both contained the actual lead compound.

Further analysis of the best library returned (SA result with objective function = 0.142) confirmed its focusing ability against the lead compound (Fig. 5). On the other hand, an analysis of molecular weight distributions revealed that the focused subset contained a large proportion of high molecular weight compounds (Fig. 6). Given that the lead itself has MW = 556.6, we should further restrict the focused design with Lipinski-like restraints.

5.3.3. Focused design with restraints

The focused design with restraints combines the previous two objectives. It attempts to select building blocks that lead to compounds similar to the target Thrombin inhibitor (Fig. 3) but also conform to the property guidelines (Table 1). We observe that random selections do not meet these objectives. Optimization of starting selections to focused and restrained subsets can be performed with SA or GA in a reliable fashion. Results from SA provided 10 libraries of which the best six contained the actual lead compound. For those four libraries that did not contain the actual lead, three out of the four R group positions (R2, R3 and R4) correctly mapped to the lead compound building blocks. Results from GA provided two libraries which, although suitable, did not contain the actual lead. For those two libraries, three out of the four R group positions (R2, R3 and R4) correctly mapped to the lead compound building blocks.

Results indicate that a design combining both objectives can be achieved and almost match results from objectives taken individually (Table 5). The focusing ability of the focused design with restraints almost matches that of the focused design alone and its penalty against restraints almost matches that of the simple restrained design. Therefore, it provides an excellent compromise where neither objective concedes any significant ground. These results confirm earlier findings by Brown et al. [9] and extend their validity to extremely large search spaces.

5.4. Sampling of virtual libraries

Sampling of the virtual libraries was a maximum of 10,000 subsets of 512 compounds for the 207 million compound Tripeptoid library (2.5%) and 10,000 subsets of 864 compounds for the 537 million compound UGI library (1.6%). Since the mutations performed in both SA and GA protocols only result in small changes in the resulting product lists, we can assume that actual sampling is much less than the above upper bounds. We conclude that satisfactory optimization of subsets can be achieved while sampling a very small fraction of the complete virtual library. Consequently, we concur with the earlier findings of Lobanov and Sheridan [12,14], that enumeration of properties for such large virtual libraries may not only be unpractical but also wasteful.

5.5. Simulated annealing versus GA

Results from GA and SA were of comparable quality for simple restrained designs. For focused designs, with or without restraints, results from GA fell in the bottom 10–20% of the SA results. Still, SA did not systematically outperform GA in our experiments. GA runs were performed in shorter run times compared to equivalent SA runs. It is possible that, given conditions for equivalent run times, results from GA would match those of SA. It is also possible that GA would outperform SA in cases where the search space is sparsely populated with solutions since SA may get stuck in a local minimum. Simulated annealing and GA can also be used in combination where GA results are further optimized with SA. We conclude that very good quality results could be obtained from both methods.

5.6. Stability of results

In order to quantify the stability of the solution provided by the optimization process, SA was performed on 10 individuals. We observed that the standard deviation from the optimization runs was considerably less than was obtained in the random runs. This result confirms the reliability of the optimization process with little dependency on the starting random selection. It is remarkable that consistent optimization can be obtained in such a large search space.

5.7. Limitations

We recognize that while considering a real combinatorial library, this analysis is purely theoretical. It would be interesting to compare the performance of subsets not purely from theoretical similarity and property considerations but also from experimental screening results.

6. Conclusion

Our results indicate that, it is not only possible but also easy to optimize library subsets against extremely large virtual libraries and search spaces. The method relies on proven approaches such as Monte Carlo, simulated annealing or GA for optimization that can be used individually or in combination. Composite optimization functions can be used to combine multiple design objectives (such as focusing and drug-likeness) and provide consistent results. Multiple design objectives can usually be met (so long as they do not conflict with each other) and provide subsets which are only marginally sub-optimal against each design objective taken individually. Finally, consistent quality subsets can be obtained while sampling only a small fraction of the virtual library and within run times that are practical for production use.

Acknowledgements

The author would like to thank Dr. Victor Lobanov at 3-Dimensional Pharmaceuticals Inc., for providing the UGI library used in the second part of this study.

References

[1] A.K. Ghose, V.N. Viswanadhan, Combinatorial library design and evaluation: principles, software tools and applications in drug discovery, Marcel Dekker, New York, NY, 2001.

[2] M.J. Valler, D. Green, Diversity screening versus focused screening in drug discovery, Drug Discov. Today 5 (2000) 286–293.

[3] S.D. Pickett, I.M. McLay, D.E. Clark, Enhancing the hit to lead properties of lead optimization libraries, J. Chem. Inf. Comput. Sci. 40 (2000) 263–272.

[4] E.J. Martin, R.W. Crichlow, Beyond mere diversity: tailoring combinatorial libraries for drug discovery, J. Comb. Chem. 1 (1999) 32–45.

[5] V.J. Gillet, P. Willett, J. Bradshaw, The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries, J. Chem. Inf. Comput. Sci. 37 (1997) 731–740.

[6] E.A. Jamois, M. Hassan, M. Waldman, Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets, J. Chem. Inf. Comput. Sci. 40 (2000) 63–70.

[7] V.J. Gillet, P. Willett, J. Bradshaw, D.V.S. Green, Selecting combinatorial libraries to optimize diversity and physical properties, J. Chem. Inf. Comput. Sci. 39 (1999) 169–177.

[8] C.H. Reynolds, A. Tropsha, L.B. Pfahler, R. Druker, S. Chakravorty, G. Ethiraj, W. Zheng, Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms, J. Chem. Inf. Comput. Sci. 41 (2001) 1470–1477.

[9] R.D. Brown, M. Hassan, M. Waldman, Combinatorial library design for diversity, cost efficiency, and drug-like character, J. Mol. Graph. Modell. 18 (2000) 427–437.

[10] V.J. Gillet, W. Khatib, P. Willett, P.J. Fleming, D.V.S. Green, Combinatorial library design using a multiobjective genetic algorithm, J. Chem. Inf. Comput. Sci. 42 (2002) 375–385.

[11] D.K. Agrafiotis, V.S. Lobanov, Ultrafast algorithm for designing focused combinatorial arrays, J. Chem. Inf. Comput. Sci. 40 (2000) 1030–1038.

[12] V.S. Lobanov, D.K. Agrafiotis, Stochastic similarity selections from large combinatorial libraries, J. Chem. Inf. Comput. Sci. 40 (2000) 460–470.

[13] R.P. Sheridan, S.K. Kearsley, Using a genetic algorithm to suggest combinatorial libraries, J. Chem. Inf. Comput. Sci. 35 (1995) 310–320.

[14] R.P. Sheridan, S.G. SanFeliciano, S.K. Kearsley, Designing targeted libraries with genetic algorithms, J. Comput.-Aided Mol. Design. 18 (2000) 320–334.

[15] R.D. Clark, D.E. Patterson, R.D. Clark, F. Soltanshahi, J.F. Blake, J.B. Matthew, Visualizing substructural fingerprints, J. Mol. Graph. Modell. 18 (2000) 404–411.

[16] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.

[17] Cerius^2, Version 4.8, Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA.

[18] R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.

[19] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.

[20] R.D. Brown, Descriptors for diversity analysis, Perspect. Drug Discov. Design 7–8 (1997) 31–49.

[21] T. Langer, R.D. Hoffmann, New principal components derived parameters describing molecular diversity of heteroaromatic residues, Quant. Struct.-Act. Relat. 17 (1998) 211–223.

[22] W.J. Egan, K.M. Merz, J.J. Baldwin, Prediction of drug absorption using multivariate statistics, J. Med. Chem. 43 (2000) 3867–3877.

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

[24] J.M. Barnard, G.M. Downs, A. Scholley-Pfab, R.D. Brown, Use of Markussh structure analysis techniques for descriptor generation and clustering of large combinatorial libraries, J. Mol. Graph. Modell. 18 (2000) 452–463.

[25] R.D. Cramer, D.E. Patterson, R.D. Clark, F. Soltanshahi, M. Lawless, Virtual compound libraries: a new approach to decision making in Molecular Discovery Research, J. Chem. Inf. Comput. Sci. 38 (1998) 1010–1023.

[26] S. Shi, Z. Peng, J. Kostrowicki, G. Paderes, A. Kuki, Efficient combinatorial filtering for desired molecular properties of reaction products, J. Mol. Graph. Modell. 18 (2000) 478–496.

[27] O. Ivanciuc, D.J. Klein, Computing Wiener-type indices for virtual combinatorial libraries generated from heteroatom containing building blocks, J. Chem. Inf. Comput. Sci. 42 (2002) 8–22.

[28] V.S. Lobanov, D.K. Agrafiotis, Combinatorial networks, J. Mol. Graph. Modell. 19 (2001) 571–578.

[29] J.S. Mason, I.M. McLay, R.A. Lewis, in: P.M. Ean, G. Nolles, C.G. Newton (Eds.), New Perspectives in Drug Design, Academic Press, London, 1995, pp. 225–253.

[30] M. Hassan, J.P. Bielawski, J.C. Hempel, M. Waldman, Optimization and visualization of molecular diversity of combinatorial libraries, Mol. Divers. 2 (1996) 64–74.

[31] J.S. Mason, S.D. Pickett, Partition-based selection, Perspect. Drug Discov. Design 78 (1997) 85–114.

[32] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci. 39 (1999) 28–35.

[33] R.N. Zuckermann, E.J. Martin, D.C. Spellmeyer, G.B. Stauber, K.R. Shoemaker, J.M. Kerr, G.M. Figliozzi, D.A. Goff, M.A. Siani, R.J. Simon, S.C. Banville, E.G. Brown, L. Wang, L.S. Richter, W.H. Moos, Discovery of nanomolar ligands for 7-transmembrane G-protein-coupled receptors from a diverse N-(substituted) glycine peptoid library, J. Med. Chem. 37 (1994) 2678–2685.