Bioorganic & Medicinal Chemistry Letters 18 (2008) 5967–5970

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Descriptor-free molecular discovery in large libraries by adaptive substituent reordering

Scott R. McAllister [a], Xiao-Jiang Feng [b,*], Peter A. DiMaggio Jr. [a], Christodoulos A. Floudas [a,*], Joshua D. Rabinowitz [b], Herschel Rabitz [b,*]

[a] Department of Chemical Engineering, Princeton University, Princeton, NJ 08544, USA
[b] Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Received 27 June 2008
Revised 8 September 2008
Accepted 10 September 2008
Available online 21 September 2008

Keywords:
Molecular discovery
QSAR
Combinatorial library
Substituent reordering
Iterative operation
Global optimization

ABSTRACT

Molecular discovery often involves identification of the best functional groups (substituents) on a scaffold. When multiple substitution sites are present, the number of possible substituent combinations can be very large. This article introduces a strategy for efficiently optimizing the substituent combinations by iterative rounds of compound sampling, substituent reordering to produce the most regular property landscape, and property estimation over the landscape. Application of this approach to a large pharmaceutical compound library demonstrates its ability to find active compounds with a threefold reduction in synthetic and assaying effort, even without knowing the molecular identity of any compound.

The discovery of new molecular entities with desired properties is a key objective in the chemical sciences. Finding such molecules can be a difficult task even with the assistance of combinatorial chemistry and high-throughput screening given the enormous number of potential candidate molecules.[1–3] To enhance the cost-effectiveness of molecular discovery, quantitative structure–activity relationship (QSAR) methods are often employed.[4–6] These methods quantify molecular properties as multi-variable functions of relevant molecular descriptors,[7] whose associated coefficients are usually determined from a training set of molecules, and the resultant parameterized functions can be utilized to predict the properties of structurally related molecules and guide laboratory synthesis. Despite their widespread use, different descriptor sets and functional forms are often needed for different classes of molecules and target properties, producing difficulties in many QSAR applications.

The general strategy of optimal substituent reordering was recently introduced to enable descriptor-free molecular discovery using minimal a priori knowledge of the molecules and the target properties.[8,9] For a molecular library under synthesis with a common scaffold, N substitution sites on the scaffold and S_i distinct substituents (functional groups) on the ith site, the technique expresses a specific molecular property y as an N-dimensional function f of the substituents bonded to the sites (Fig. 1). The collection of all potential library compounds and their property values then form an $N+1$-dimensional property landscape. The nature of the substituents is captured by assigning each of the S_i substituents on the ith site a random, but distinct, integer value $X_i \in [1,S_i]$. As a result, the structure and property function value of each molecule is uniquely associated with the integer assignment for each substituent on each site. Note that this indexing method does not require any traditional molecular descriptors.

Based on this substituent indexing scheme, molecular discovery is performed by (1) randomly sampling the variables (i.e., synthesizing a small random subset of the potential library molecules) and measuring the targeted property value for each molecule (Fig. 1, Step (1.a)), and (2) estimating/interpolating over the property landscape to find molecules with desirable property values. From an initial random integer assignment, however, the $N+1$-dimensional property landscape will most likely be highly irregular (Fig. 1, Step (1.b)) and provide no estimation/interpolation capability for finding the desired molecule(s). In order for the technique to have predictive power, the critical operation in Step (2.a) identifies the optimal substituent ordering (i.e., the optimal integer assignment for each substituent) on each site that results in a property landscape with regular structure. Property estimation/interpolation over the landscape can then be readily implemented (Fig. 1, Step (2.b)).[8,9]

When the size of the molecular library is large, the most efficient implementation of the reordering technique involves iterative rounds of compound synthesis and data reordering, starting

[*] Corresponding authors.
E-mail address: xfeng@princeton.edu (X.-J. Feng).

0960-894X/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2008.09.068

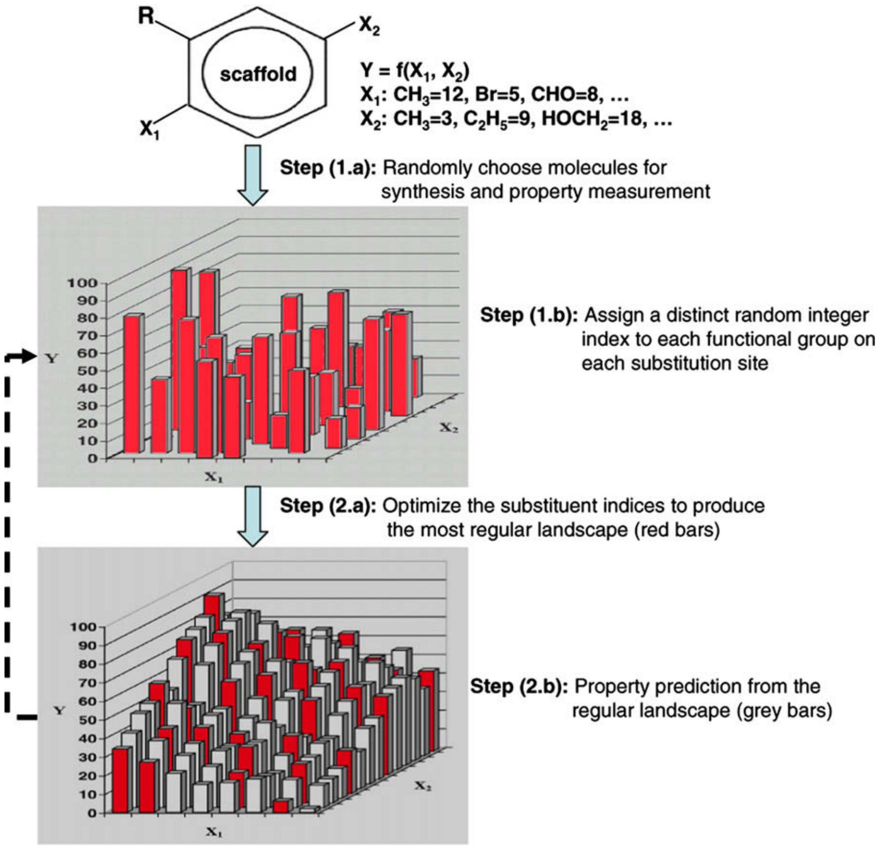
R
X2
X1
scaffold
Y = f(X1, X2)
X1: CH3=12, Br=5, CHO=8, ...
X2: CH3=3, C2H5=9, HOCH2=18, ...
Step (1.a): Randomly choose molecules for synthesis and property measurement
Step (1.b): Assign a distinct random integer index to each functional group on each substitution site
Step (2.a): Optimize the substituent indices to produce the most regular landscape (red bars)
Step (2.b): Property prediction from the regular landscape (grey bars)
Y
X1
X2
0
10
20
30
40
50
60
70
80
90
100

Figure 1. General operation of the adaptive substituent reordering technique. As an example, for a library with a common scaffold and two substitution sites, the property y of any molecule is represented as a two-variable function f (with a priori unknown form) of the substituents (X_1 and X_2) on the two sites. Each substituent on each site is represented by a distinct integer, hence the property of each compound is uniquely determined by the integer assignments on both sites. Step (1.a): Molecular discovery begins with initial synthesis and property measurement of a random subset of the library, resulting in (most likely) an irregular property landscape (with little predictive capability) from a random ordering of the substituents (i.e., the collection of random integer assignments for each substituent, Step(1.b)). Step (2.a): Suitable optimization algorithms are employed to identify the substituent orderings that generate the most regular landscape. Step (2.b): The resultant smooth landscape can be used to make property predictions of the unsynthesized compounds (the gray bars). The process operates iteratively (the dashed line) until desired property prediction is obtained. The 3-D bar graphs are made from the data in Ref. 8 for a co-polymer library with y being the glass transition temperature; the gray bars are laboratory data placed at locations dicated by the optimal orderings of the substituents associated with the red bars.

with a minimal sampling of the library space (Fig. 1).[9] In each iteration, Step (2.b) provides an estimation/prediction of the library domain most likely to be enriched in promising compounds. Compound synthesis and property assaying can then be guided by this estimate, providing additional data to further enhance the reliability of substituent reordering and property prediction. This adaptive operation can be viewed as a process of attaining enhanced resolution and regularity over the landscape with each iteration until the full property landscape is revealed to the desired degree.

Previous applications of the reordering strategy were to small compound libraries[8,9] and without iterative operation. This article provides the first illustration of the technique on a large pharmaceutical compound library utilizing the adaptive reordering procedure, where the measured property is percent inhibition of a protein function.[12] All of the compounds have a common scaffold with $N = 2$ ($S_1 = 151$ and $S_2 = 93$). Of all the 14,043 potential library compounds, data is available for 4110 (29%) (Fig. 2(a)). The goal is to identify the high-inhibition compounds over the whole library space from sampling a small number of compounds.

In this work, the regularity of the property landscape is quantified by a global pairwise difference measure

$$Q = \sum_{m=1}^{N}\left[\sum_{\substack{n=1\\ n\neq m}}^{N}\sum_{j=1}^{S_n}\sum_{i=1}^{S_m}\sum_{i'=1}^{S_m}\left(\frac{1}{w^m}\cdot\frac{S_m - d^m_{i,i'}}{S_m - 1}\right)\cdot\left(a^{m,n}_{i,j} - a^{m,n}_{i',j}\right)^2\right],$$

where N is the number of substitution sites, S_m and S_n are the total number of substituents on the mth and the nth site, respectively, $a^{m,n}_{i,j}$ is the measured property value of a sampled molecule whose substituents are assigned to integer i at the mth site and integer j at the nth site, $a^{m,n}_{i',j}$ is the property value of another compound that differs only in the integer assignment (i') at the mth site, $d^m_{i,i'} = |i - i'|$ is the distance between these two compounds at the mth site, and w^m is the number of compound pairs where both $a^{m,n}_{i,j}$ and $a^{m,n}_{i',j}$ values are available from synthesis and property assaying over i and i'. With this form, minimization of Q tends to place together compounds with similar property values, resulting in the most regular property landscape(s). Other appropriate forms of Q can be used as well.[8,9] The minimization of Q can be achieved by several deterministic[10] and stochastic[11] optimization algorithms. The results presented in this paper are from the deterministic method.

The effectiveness of the reordering technique without iterative operation was first evaluated. We randomly selected 2055 compounds (i.e., 50% of the available data and 15% of the whole library) and determined the optimal substituent orderings (Fig. 2(b)) on both sites that resulted in the most regular inhibition landscape. Figure 2(c) applies the identified best ordering in Figure 2(b) to all the available compounds in the library. In both figures, there is a significant clustering of the high-inhibition compounds (red) in the upper left corner of the property landscape.

The capability of the reordering technique is evident from consideration of Figure 2(b) and (c). One can view Figure 2(b) as ‘pre-

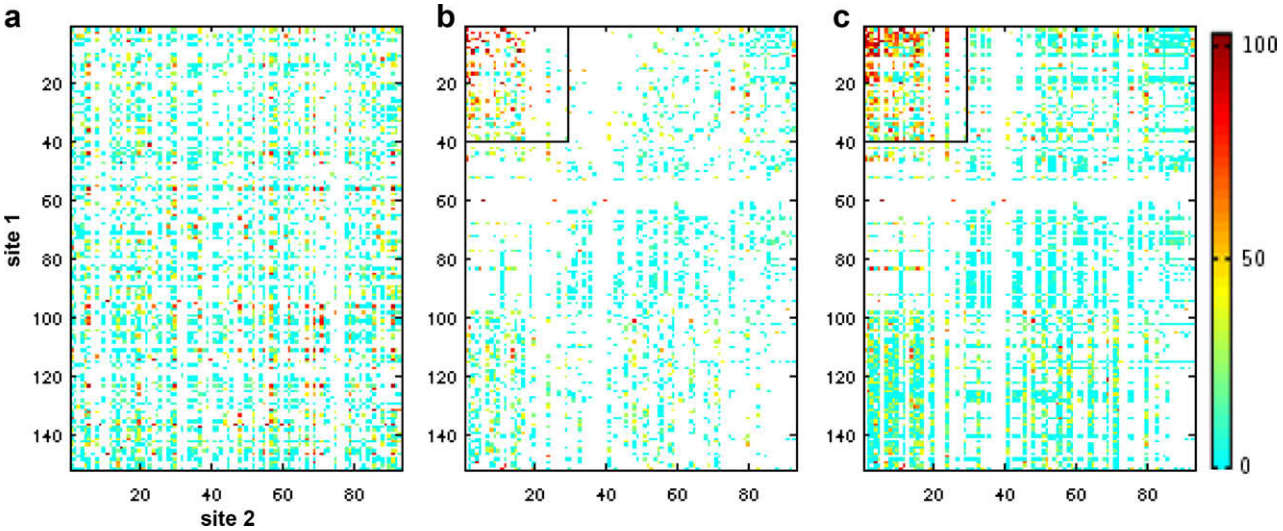

Figure 2. Heat maps of inhibitor efficacies[12] prior to, (a), and following, (b) and (c), substituent ordering optimization. Percentage inhibition is color coded (see key at right side of figure) with white indicating unsynthesized compounds, dark red the best inhibitors, and light blue the worst inhibitors. Each square in the matrix reflects a specific compound with different substituents at the two substitution sites. There are 151 substituents on site 1 and 93 on site 2. Data is available for 29% of the full library.[12] Relative data error is estimated to be ~15% from repetitive measurements of some compounds; the mean inhibition values are used in the analysis. (a) The property landscape with a random substituent ordering. (b) The optimal substituent ordering obtained by using a random subset of 15% of the library space (i.e., 50% of the available data in (a)). (c) The library landscape containing all available data, using the optimal ordering shown in (b). Without knowing the identity of any compound, the algorithm predicts that unsynthesized compounds located in the upper left corner of Figure 2(b) should be enriched in effective inhibitors. This is confirmed by the remaining data (Fig. 2(c)). Performing synthesis in the indicated boxed region of Figure 2(b) is ~50 times more effective than random synthesis for finding effective inhibitors (i.e., compounds with >70% inhibition).

dicting' that further synthesis in the upper left corner would be more effective than anywhere else in the property landscape, and utilization of the remaining data in Figure 2(c) confirms this prediction. In order to quantitatively assess the algorithm's prediction capability, we conservatively select a box of size 40 × 30 at the upper left domain of the library in Figure 2(b), which includes 52 out of 64 high-inhibition compounds (i.e., those with inhibition value greater than 70%). When all the remaining compounds are placed in the landscape (Fig. 2(c)), this box includes 98 out of 121 high-inhibition compounds, corresponding to synthesis in this domain being nearly 50 times more effective than if it were performed elsewhere over the landscape.

Following the above test, the more efficient iterative reordering procedure was examined, as iterative operation is most likely to occur in practical applications. This test was performed by first randomly selecting a small initial subset of the available compounds. The optimal substituent ordering corresponding to this subset was then determined, and linear interpolation was employed to estimate the 50 unsampled compounds with the highest inhibition values. The laboratory data for these compounds were then combined with that of the initial subset to improve the reliability of substituent reordering and property estimation. This iterative process was carried out until the data for 3000 compounds were utilized.

Figure 3 shows the number of desired compounds (i.e., those with >70% inhibition) discovered in the iterative process. Each curve corresponds to a particular number of initially sampled compounds. All curves exhibit a sharp slope in the beginning, indicating that most of the desired compounds may be discovered at the early stage of the adaptive operation. Interestingly, starting from a smaller initial sampling generally results in the most rapid discovery of similar number of high-inhibition compounds. For example, if the iterative process terminates when 1/2 of the available compounds above 70% inhibition are found, then the number of compounds needed to be synthesized is ~650 for an initial sampling of 5% of the available compounds, compared with >1,000 for the initial sampling of 25% (Fig. 3). For the non-iterative method, >2000 samples are required (starting point of the purple curve in Fig. 3). Thus, a factor of ~3 savings in synthesis and property assaying effort arises from the iterative operation by starting from an absolute minimal library, which can make a significant difference in many circumstances.

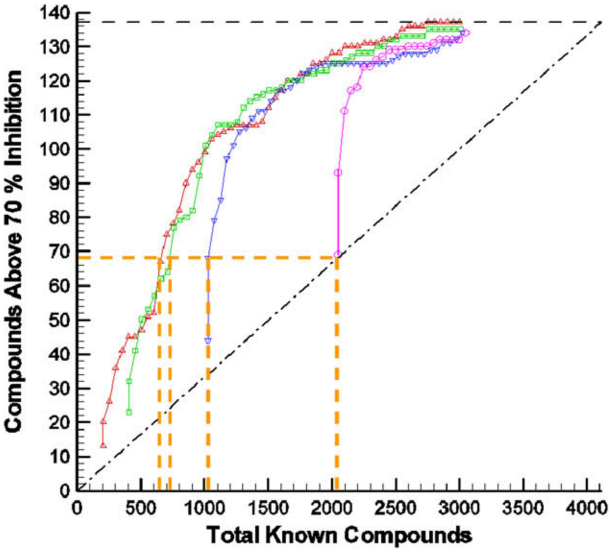

Figure 3. The number of desired compounds (i.e., those with >70% inhibition) discovered through the iterative reordering procedure. The red, green, blue, and purple curves correspond to initial samplings of 5%, 10%, 25%, and 50% of the available data, respectively. Based on the predicted property values in each iteration, 50 new compounds are then added. The horizontal black dashed line at 137 shows the total number of desired compounds. The orange dashed lines show the number of compounds required for synthesis to discover 1/2 of the desired compounds. Significantly, the algorithm is the most efficient by starting with minimal initial sampling; only 650 compounds out of a total library of ~14,000 potential compounds suffices when operating in this fashion.

The reordering technique has an important feature of not requiring any traditional molecular descriptors. When the molecular scaffold and the substituents are chosen, the indexing scheme always provides a complete and unambiguous set of 'canonical descriptors' to represent functionally related molecules. One does not even need to know the structure of the molecules[12] to apply the reordering technique, as all relevant information lies in the encoded relationship between the substituent indices and the property measurements. In addition, knowledge of the explicit form of the property–structure relationship function $y = f(X_1, X_2, \ldots, X_N)$ is not necessary for the reordering operation. Due to these features, the reordering technique may be readily applied to a diverse array of molecular discovery problems, regardless of molecular types (e.g., from small molecules to peptides) or target properties (e.g., from electronic properties to biological attributes). The reordering technique can be implemented to any case where (a) the library molecules can be identified by site and substituent indices (i and X_i, respectively) and (b) property data y is available for an adequate subset of the molecules. In cases where the library molecules contain more than one common scaffold or it is hard to define a common scaffold, different means of uniquely encoding the library molecules are required. This topic is a subject of ongoing research.

The reordering strategy and traditional QSAR methods should not be viewed as competing techniques. Substituent reordering is inherently an interpolation method, hence unlike QSAR, it cannot extrapolate over substituents that are unsampled across all related substitution sites. In addition, the reordering strategy does not directly provide structure–property relationships. However, the strategy can enhance the effectiveness of QSAR methods by identifying functionally similar substituents (with respect to the molecular property of interest), which locate adjacently on a particular substitution site in the optimal orderings. In this case the nature of the compounds and property assay must be known. However, the example in this paper shows that the substituent reordering technique can successfully function even without this knowledge. Considering their complementary advantages, suitable integration of the reordering technique and standard QSAR methods is expected to synergistically benefit each other and enable more efficient molecular discovery and more reliable understanding of structure–property relationships.

In summary, the adaptive reordering technique provides a practical and easy-to-use means for a broad variety of molecular discovery tasks. On the laboratory side, one only needs to randomly sample a small subset of the target compound library, assign distinct random integers to the functional groups on each substitution site, and measure the target property of the subset. The data is then fed to the reordering algorithm(s), which will generate property predictions and provide suggestions on further laboratory synthesis and assaying. We are building an easily understandable graphical user interface to the core algorithms for the convenience of the end users.

Acknowledgments

The authors are grateful to Dr. C.-H. Tseng of Pfizer Inc. for providing the data utilized in this work. The authors acknowledge support from the NSF and the US EPA (GAD R 832721-010).

References and notes

1. Ng, R. Doe *Drugs—from Discovery to Approval*; Wiley Liss: New Jersey, 2006.
2. Bannwarth, W.; Hinzen, B.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Combinatorial Chemistry: from Theory to Application (Methods and Principles in Medicinal Chemistry)*; Wiley-VCH: New Jersey, 2006.
3. Huser, J.; Mannhold, R.; Kubinyi, H.; Folkers, G. *High-throughput Screening in Drug Discovery (Methods and Principles in Medicinal Chemistry)*; Wiley-VCH: New Jersey, 2006.
4. Hansch, C.; Leo, A. *Exploring QSAR—Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
5. Leach, A. *Molecular Modeling: Principles and Applications*; Prentice Hall, 2001.
6. *Predictive Toxicology*; Hemla, C., Ed.; CRC, 2005.
7. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, 2000.
8. Shenvi, N.; Geremia, J.; Rabitz, H. *J. Phys. Chem. A* **2003**, *107*, 2066.
9. Liang, F.; Feng, X.; Lowry, M.; Rabitz, H. *J. Phys. Chem. B* **2005**, *109*, 5842.
10. Floudas, C. A. *Nonlinear and Mixed Integer Optimization*; Oxford University Press: New York, 1995.
11. Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley, 1989.
12. The compounds in the library are proprietary to Pfizer Inc. The data was provided to the authors in an encoded form with each compound labeled by two integers, representing the functional groups at the two substitution sites, respectively. The fact that the specific molecules involved (both drugs and targets) were not revealed by Pfizer enables (a) an unbiased evaluation of the reordering technique and (b) an excellent demonstration of the capabilities of the method.