Identifying Promising Compounds in Drug Discovery: Genetic Algorithms and Some New Statistical Techniques

Abhyuday Mandal*

Department of Statistics, University of Georgia, Athens, Georgia 30602-1952

Kjell Johnson*

Pfizer Global Research and Development, Michigan Laboratories, Ann Arbor, Michigan 48105

C. F. Jeff Wu

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205

Dirk Bornemeier

Pfizer Global Research and Development, Michigan Laboratories, Ann Arbor, Michigan 48105

Received December 14, 2006

Throughout the drug discovery process, discovery teams are compelled to use statistics for making decisions using data from a variety of inputs. For instance, teams are asked to prioritize compounds for subsequent stages of the drug discovery process, given results from multiple screens. To assist in the prioritization process, we propose a desirability function to account for a priori scientific knowledge; compounds can then be prioritized based on their desirability scores. In addition to identifying existing desirable compounds, teams often use prior knowledge to suggest new, potentially promising compounds to be created in the laboratory. Because the chemistry space to search can be dauntingly large, we propose the sequential elimination of level combinations (SELC) method for identifying new optimal compounds. We illustrate this method on a combinatorial chemistry example.

INTRODUCTION

Historically in the drug discovery process, work has primarily centered on finding chemical entities that are effective against a particular disease or condition. Effective chemical entities are often found through an iterative synthesis and screening process. Upon finding an effective compound or series of compounds, work then focuses on tweaking the molecule(s) to eliminate potential negative effects and to improve the molecule's ability to interact with the body. Historically, this two-stage process has found numerous blockbuster pharmaceutical entities.

Over the past decade, great advancements have been made in the automation of in vitro biological screening.[1] This improving technology enables companies to quickly test more compounds at lower concentrations across a number of screens. Thus, companies now have a large amount of information about more chemical entities earlier in the drug discovery process.

With the increase in information, drug discovery teams have begun to turn from the two-stage optimization process to a process that *simultaneously* optimizes over a variety of efficacy, pharmacokinetic/dynamic, and safety endpoints. Hence, instead of focusing only on information about the effectiveness of the chemical entities, the team can now also focus on other desirable endpoints.

Given this vast amount of data currently available, the process of prioritizing compounds for follow-up can be extremely difficult. Often, no individual compound provides the optimal value for each endpoint under consideration. Instead, many compounds are near optimal for one or more endpoints. To increase the complexity of the problem, discovery teams rarely place an equal weight of decision on each endpoint.

Many simple, univariate approaches can be taken for solving the multiple endpoint problem. However, the univariate approaches do not consider the potential interdependencies between endpoints or their weighting of importance.

In addition to improvements in screening capacity of compounds, technologies have been developed to explore and synthesize vast numbers of chemical entities. This technology, known as combinatorial chemistry, has been widely applied in the pharmaceutical industry and is gaining interest in other areas of chemical manufacturing.[2,3] In general, combinatorial chemistry identifies molecules that can be easily joined together and employs robotics to physically make each molecular combination. Depending on the initial number of molecules, the number of combinations can be extremely large. For example, consider a core molecule onto which various reagents can be theoretically added to three locations. If 100 reagents can be added at each location on the core, then 1 million products can potentially be synthesized. In the pharmaceutical industry, combinatorial chemistry has been used to enhance the

* Corresponding author e-mail: amandal@stat.uga.edu (A.M.) and kjell.johnson@pfizer.com (K.J.).

10.1021/ci600556v CCC: $37.00
Published on Web 04/11/2007

Table 1. Desirable Ranges for Compound Prioritization

end points	desired type	acceptable range
Y_1	smaller the better	$Y_1 < 10$
Y_2	larger the better	$Y_2 > 500$
Y_3	smaller the better	$Y_3 < 10$
Y_4	smaller the better	$Y_4 < 60$
Y_5	larger the better	$Y_5 > 1$
Y_6	nominal the best	$20 < Y_6 < 80$

diversity of compound libraries, to explore specific regions of chemical space (i.e., focused library design), and to optimize one or more pharmaceutical endpoints such as target efficacy or ADMET (absorption, distribution, metabolism, excretion, toxicology) properties.[4] While it is theoretically possible to make a large number of chemical combinations, it is generally not possible to follow up on each newly synthesized entity. An alternative approach to synthesizing all possible molecular combinations is to computationally create and evaluate the entire library using structure-based models. (For this purpose, specialized software uses "black box" type functions.) Then, a subset of promising compounds is selected for synthesis. For the purpose of optimization of pharmaceutical endpoints, the proposed sequential elimination of level combinations (SELC) method can be employed to efficiently find optimal molecules, as will be demonstrated in the proof of concept example.

In the first part of this paper, we explore the use of the desirability function for compound prioritization. While this function is simple to compute and interpret, it allows the user to incorporate a priori scientific knowledge or endpoint prioritization. In the second part, we present the SELC optimization technique to assist in the creation of new compounds in combinatorial chemistry.

TECHNIQUE 1: COMPOUND PRIORITIZATION

In our first example, compounds are to be prioritized for subsequent stages of the drug discovery process, given results from multiple screens. For example, consider a project for which we have measured six quantities for each compound: Y_1, ..., Y_6. Larger values are better ("larger the better") for two qualities, smaller values are better ("smaller the better") for three qualities, and a target value is best ("nominal the best") for one quality. More specifically, a compound will be good if Y_1, Y_3, and Y_4 are small, Y_2 and Y_5 are large, and Y_6 is close to 42.5 (Table 1).

Unfortunately, no compounds for this project meet all of the desirable characteristics. Given this situation, how can we prioritize compounds for follow-up? Several statistical methods are available for treating this kind of multiple-response problem. Myers and Montgomery[5] used a graphical method, which has clear disadvantages. A more general approach is to formulate it as a constrained optimization problem, where one of the responses is selected as the objective function and the other responses are treated as constraints. This technique is also not recommended as the choice of objective function can be debatable. For details, see ref 6.

Desirability Score. A third technique is to combine the information into one numeric score which can be used to prioritize the compounds. The techniques available for combining multiple-response models into a single scalar include distance functions,[7] squared error loss functions,[8,9] and desirability functions.[10−12] The desirability methods are easy to understand and implement, are available in software, and provide flexibility in weighting individual responses. This approach consists of transforming the individual response functions each into "desirability scores" based on the particular goal for that response. Individual desirabilities $d_i(\hat{y}_i)$, $i = 1,...,6$, map response values to unitless utilities bounded by $0 < d_i(\hat{y}_i) < 1$, where a higher value of d_i indicates that response value $\hat{y}_i$ is more desirable; hence it is termed as a "desirability score". The individual desirability scores, d_i, are combined into one overall desirability score for the element using either a multiplicative or additive model. A common approach is to define the overall desirability as the geometric mean of individual desirability d_i's, $i = 1,..., m = 6$, where

$$d = \{d_1 d_2 ... d_m\}^{1/m}$$

If all the endpoints are not equally important, then the definition of overall desirability function can be extended to

$$d = d_1^{w_2} d_2^{w_2} ... d_m^{w_m}$$

reflect the possible difference in the importance of the different responses, where the weights w_i satisfy $0 < w_i < 1$ and $w_1 + w_2 + ... + w_m = 1$.

Although several forms have been proposed for $d_i(\hat{y}_i)$, the most commonly adopted are those of Derringer and Suich.[11] It is a two-sided problem for the *nominal-the-best* case, where it is ideal for a compound to have the Y_6 score as close as possible to a target value denoted by t. Apart from the target value t, there is a lower value L and an upper value U such that the product is considered unacceptable if $Y_6 < L$ or $Y_6 > U$. The desirability function is then defined as

$$d_6 = \begin{cases} \left|\dfrac{\hat{y} - L}{t - L}\right|^{\alpha_1}, & L \le \hat{y} \le t \\ \left|\dfrac{\hat{y} - U}{t - U}\right|^{\alpha_2}, & t \le \hat{y} \le U \end{cases}$$

with $d_6 = 0$ for $\hat{y} < L$ or $\hat{y} > U$. The choice of α_1 and α_2 is more subjective than the choice of L and U, and these quantities define the penalty we pay for moving away from t.

Next we consider the *smaller-the-better* problem with a being the smallest possible value for the response y. For this example, Y_1, Y_3, and Y_4 are denoted by y, and the corresponding observed value is denoted by $\hat{y}$. Treat a as the target value and choose U to be a value above which the product is considered to be unacceptable. For example, U for Y_1 is 10 and the corresponding a can be taken to be 0 because the response cannot be negative. Then we can choose the right half of the d_i function as the desirability function, that is, let

$$d_i = \left|\frac{\hat{y} - U}{a - U}\right|^{\alpha}, \quad a \le \hat{y} \le U$$

with $d_i = 0$ for $\hat{y} > U$.

For the *larger-the-better* problem, there is no fixed ideal target. Suppose that the scientist can choose a value L below which the compound is considered to be unacceptable, and

a value U above which it is considered to be nearly perfect. Then the desirability function can be defined as

$$d_i = \left|\frac{\hat{y} - L}{U - L}\right|^{\alpha}, L \le \hat{y} \le U$$

with $d_i = 0$ for $\hat{y} < L$ and $d_i = 1$ for $\hat{y} > U$. For this example U for Y_2 and Y_5 are taken to be 1000 and 5, respectively.

While the desirability function is flexible, its original form is biased toward individual undesirable characteristics. That is, the overall desirability function can be driven to zero if any $\hat{y}_i$ is either infeasible or otherwise undesirable. This strict penalization is not desirable because some compounds may perform poorly for only one endpoint, even though all other endpoints may be highly desirable. Additive desirability functions, such as the arithmetic mean discussed in Kros and Mastrangelo,[13] incorporate a weak penalty when a constraint has been violated. However, additive methods for combining the individual desirability scores can result in unacceptable compounds having higher desirability values than acceptable ones. In order to overcome this, Ortiz et al.[14] proposed unconstrained multiplicative desirability function which does not allow unacceptable compounds to have higher overall desirability values than acceptable ones. This method involves incorporating the constraints directly into the overall desirability function via penalties. The overall desirability function, $D^* = d - p$, incorporates penalties through p, which is proportional to the square of the constraint violation. The overall penalty function p is also a combined function of the individual fitted responses, reflecting the overall severity of the infeasibility. The overall penalty function is

$$p = [\{p_1 p_2 ... p_m\}^{1/m} - c]^2$$

where c is a relatively small constant to force $p_i > 0$. Smaller or larger values of c can be used without loss of generality. The corresponding individual penalties $p_i(\hat{y}_i)$ are proposed as follows. For *nominal-the-best* case

$$p_6 = \begin{cases} c + \left|\frac{\hat{y} - L}{t - L}\right|^{\beta_1}, & 0 \le \hat{y} \le L \\ c, & L \le \hat{y} \le U \\ c + \left|\frac{\hat{y} - U}{t - U}\right|^{\beta_2}, & U \le \hat{y} \end{cases}$$

for *larger-the-better*

$$p_i = \begin{cases} c + \left|\frac{\hat{y} - L}{U - L}\right|^{\beta}, & \hat{y} \le L \\ c, & \hat{y} > L \end{cases}$$

and for *smaller-the-better*

$$p_i = \begin{cases} c, & \hat{y} \le U \\ c + \left|\frac{\hat{y} - U}{a - U}\right|^{\beta}, & \hat{y} > L \end{cases}$$

Incorporating this overall penalty function into a combined fitted response metric, the proposed overall unweighted desirability function becomes

$$D^* = \{d_1 d_2 ... d_m\}^{1/m} - [\{p_1 p_2 ... p_m\}^{1/m} - c]^2$$

The penalty modification improves the flexibility of the desirability function by preventing poor individual desirabilities from dominating the product. Because of the flexibility of this function, we suggest that the user determine a defendable range of parameter settings prior to computing the desirability scores in order to avoid biasing the resulting scores.

ILLUSTRATION

For our example data, a is taken to be zero and the α's and β's are taken to be unity. The d values of the compounds corresponding to different values of c are plotted in Figure 1. The y-axis gives the desirability scores, and the important compounds are marked on the plots with their index numbers. It can be easily seen that, with higher values of c, worse compounds are well separated. Compound numbers 8, 35, 65, 105, 120, 123, 169, 188, and 206 turn out to be desirable, whereas compounds 52 (also 51 and 53) are undesirable.

For this analysis, α's and β's are taken to be 1, which leads to a linear desirability function. However, these choices are often not the best for all problems. For example, one might use a small α value if the response does not have to be very close to the target t. On the other hand, a large α value would imply the importance of being close to t. For nominal-the-best case, if the penalties for being above or below the target are very different, this difference can be reflected by choosing different values for α_1 and α_2.

To illustrate the effect of different choices of α and β, we have selected an undesirable compound (52) and a highly desirable compound (188). For each of these compounds we have computed the desirability score across a range of α and β (keeping each parameter the same for each characteristic, $Y_1 - Y_6$) for $c = 0.01$ and 0.00001. For the undesirable compound, changes in α have no affect on the overall desirability score (Figure 2). However, the choice of both β and c have an affect on the overall score: as each increase, the desirability score is dampened toward zero. This makes sense because both β and c are constructed to minimize the effect of any individual undesirable characteristic. For a highly desirable compound, changes in β and c have no affect on the overall desirability score, whereas α has a significant effect on the score (Figure 2). As α increases, the desirability score decreases, which must occur because the individual desirabilities are scaled between 0 and 1.

TECHNIQUE 2: OPTIMIZATION IN COMBINATORIAL CHEMISTRY

A simple combinatorial chemistry problem involves connecting reagents to each of several locations along a scaffold. Usually many potential reagents can be added at each location on a scaffold, and as the number of locations and number of reagents increase, the number of compounds to create increases exponentially. Because of the vast number of potential compounds, the entire combinatorial library is rarely created in practice. Instead, scientific knowledge about the project and the reactions are used to select a subset of reagents for each location on the scaffold. After creating the first library, subsequent libraries are created that use the knowledge gained from the first library.

Indeed, several different optimization techniques could be applied to this problem. Generally, optimization techniques utilize either a systematic search (e.g., response surface or

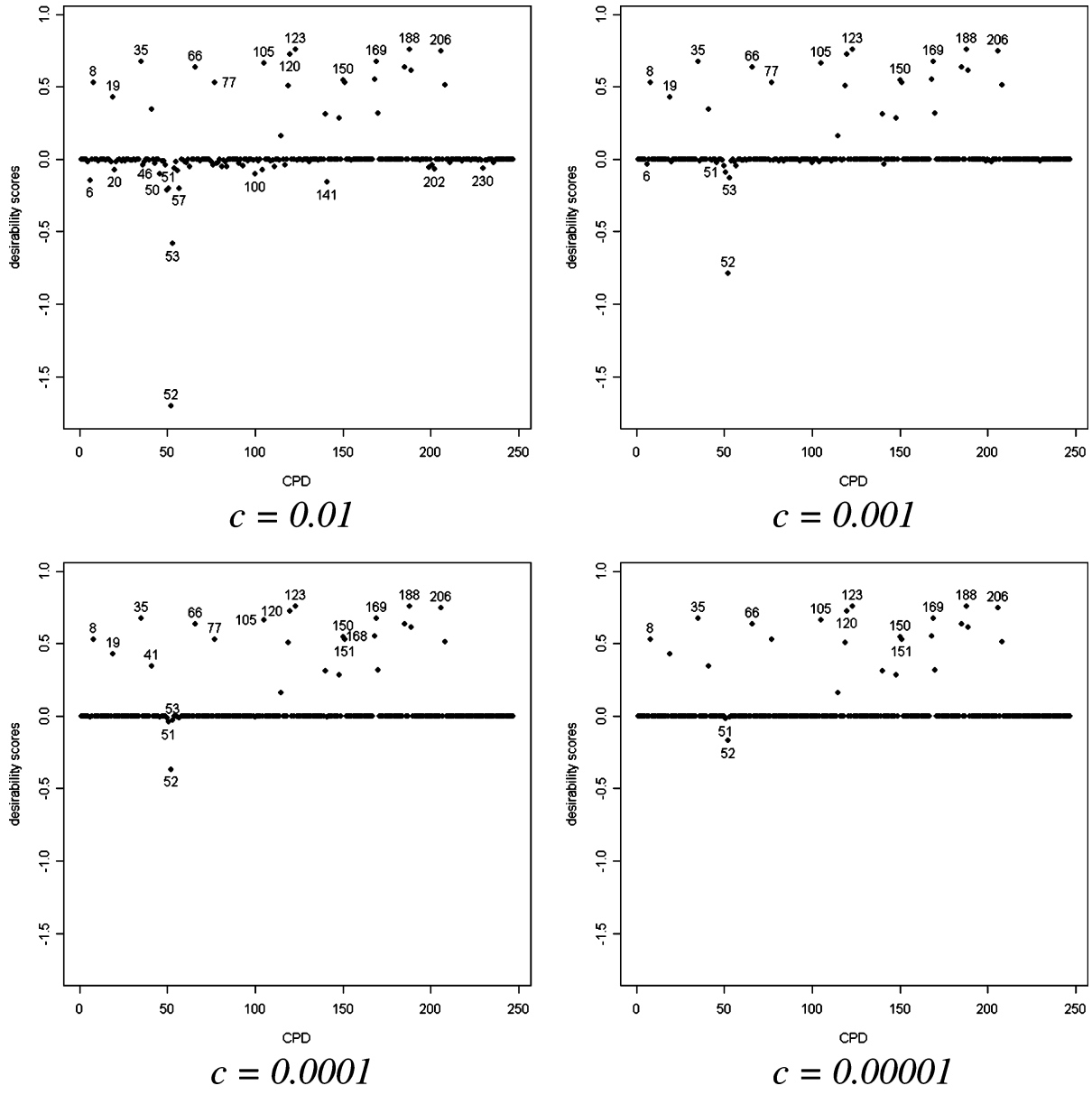

Figure 1. Desirability scores for each compound for different values of *c*.

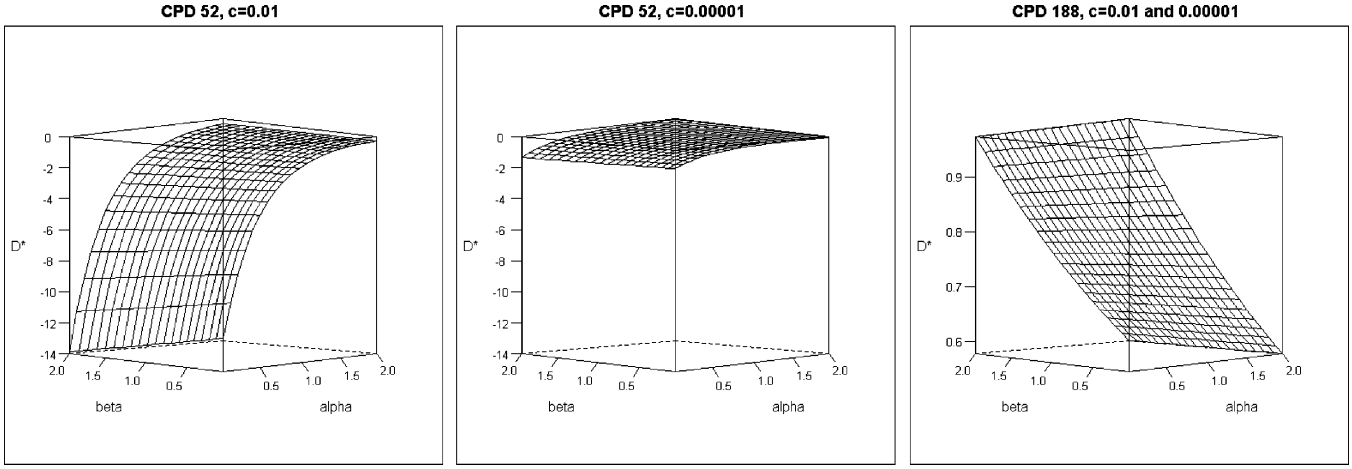

Figure 2. The combined effects of each parameter on the desirability function for compounds 52 and 188.

Newton−Raphson methods) or a search with a random component (e.g., genetic algorithms or simulated annealing). Systematic search techniques are known to be best for finding optimal values from a smooth response surface. Alternatively, search techniques with a random component are best for finding global optimums on a response surface where there are local optimums. For many problems, like those in combinatorial chemistry, the response surface is likely to

have a variety of shapes, where some parts of the surface are smooth, other parts are filled with local optimums, and others have unexpected extreme peaks of activity. Hence, to find optimums in this type of space, the technique needs to have both systematic and random components. Furthermore, the technique needs to incorporate expert opinion to guide the search process; standard optimization techniques do not incorporate expert opinion, which is the chief reason why they are not used for problems involving complex response surfaces. As an alterative to traditional optimization techniques, we suggest using the sequential elimination of level combinations (SELC) method.[15] Specifically, SELC uses concepts of experimental design and genetic algorithms in addition to expert opinion to find optimal candidates from a very large pool of potential candidates. Mandal et al.[15] show that for problems with complicated response surfaces, the SELC finds optimums more efficiently than either experimental design or genetic algorithms alone. Before describing the novel approach of employing SELC in combinatorial chemistry, we briefly review GAs.[16]

GAs are stochastic optimization tools that work on "Darwinian" models of population biology and are capable of obtaining near-optimal solutions for multivariate functions without the usual mathematical requirements of strict continuity, differentiability, convexity, or other properties. The algorithm attempts to mimic the natural evolution of a population by allowing solutions to reproduce, creating new solutions, and to compete for survival. It begins by choosing a large number of candidate solutions which propagate themselves through a "fitness criteria" and are changed by the application of well-developed genetic operators. The idea of GAs is to get "better candidates" using "good candidates", and the algorithm process is as follows:

1. **Solution representation:** For problems that require real number solutions, a simple binary representation is used where unique binary integers are mapped onto some range of the real line. Each bit is called a *gene*, and this binary representation is called *chromosome*.

Once a representation is chosen, the GA proceeds as follows. A large initial population of random candidate solutions is generated; these are then continually transformed following steps 2 and 3.

2. *Select* the best and *eliminate* the worst solution on the basis of a fitness criterion (e.g., higher the better for a maximization problem) to generate the next population of candidate solutions.

3. *Reproduce* to transform the population into another set of solutions by applying the genetic operations of "crossover" and "mutation". (a) Crossover: A pair of binary integers (chromosomes) is split at a random position, and the head of one is combined with the tail of other and vice versa. (b) Mutation: The state (0 or 1) of a randomly chosen bit is changed. This helps the search avoid being trapped into local optima.

4. *Repeat* steps 2 and 3 until some convergence criterion is met, or some fixed number of generations has passed.

This algorithm has been shown to converge by Holland,[17] who first proposed this procedure in its most abstract form and discussed it in relation to adaptive and nonlinear systems.

Now we explain the SELC method, in the context of a combinatorial chemistry example. Consider the problem of maximizing the objective function $y(x) = f(x_1,x_2,...,x_n)$ whose analytic form is unknown and which is very complex in nature. However, for a given value of $x_1,x_2,...,x_n$, $y(x)$ can be evaluated. These evaluations are expensive, and hence the total number of possible evaluations is limited by the available resources. For $i=1,..., n$, x_i can take discrete values denoted by 1,2,..., s_i. Hence there are $s = \Pi_{i=1}^{n} s_i$ possible candidates, and the challenge is to find the optimal candidate which maximizes y with a limited number of evaluations of the unknown function f.

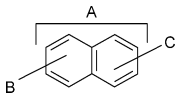

Figure 3. Combinatorial scaffold. In this example, there are 5 possible substructures at position A, 34 at position B, and 241 at position C.

Consider the combinatorial chemistry example presented in Figure 3, where three positions require additions. These three positions of the core molecule are denoted by x_1,x_2,x_3, and each x_i is called a "factor". The different reagents to be added to those positions are called "levels" of the factors, which means factor x_i has s_i levels denoted by 1,2,..., s_i. If $s_1 = s_2 = s_3 = 20$, there are $20^3 = 8000$ possible compounds, among which, say, we are constrained to creating only a fraction in the laboratory. The response function y is obtained in a follow-up experiment after the synthesis of the new compound.

The SELC Algorithm. In the absence of prior knowledge about the experiment, the SELC is initialized with an orthogonal experimental design.[18] Ideally roughly one-fourth of the available resources should be used to conduct the initial experiment with the remaining resources used for follow-up runs (i.e., new promising compounds). The data from this initial experiment are then used to identify runs that are and are not optimal via a "fitness" measure (i.e., value of y). Runs that are optimal are used to generate subsequent runs, while the runs that are not optimal are placed into a *forbidden array*. The purpose of the forbidden array is to prevent potentially poor compounds from being synthesized in subsequent experiments and is defined by its *strength* and *order*. (Based on scientific knowledge, chemists can often identify runs which are unlikely to yield desirable results prior to the initial experiment. In this case, this information can be placed into the forbidden array before the initial experiment.) The number of runs selected for the forbidden array for each stage defines the array's strength, while the number of level combinations prohibited from appearing in future runs defines the array's order. For example, a forbidden array of order k means that any combinations of k or more levels from any run in the forbidden array will be eliminated from consideration of being created in future experiments. Thus, as the order decreases, the number of forbidden design points increases.

After constructing the forbidden array, SELC starts searching for better level settings using GAs. Typically the two best runs are chosen, with probability proportional to the "fitness", i.e. the value of y, to generate potentially better candidates. These two runs, called "parents", are split at a random position, and the top fragment of one is combined with the bottom of the other and vice versa to produce two new runs called "offspring" (i.e., crossover). The next step of generating new runs is called "mutation", where a factor

Table 2. New Substructure Space Explored by Each Historical Iteration

iteration	no. of compds created	A	B	C	A × B	A × C	B × C
1	2114	5	26	164	75	286	1168
2	208	0	7	9	17	18	102
3	128	0	1	41	1	41	120
4	33	0	0	27	0	27	32
overall	2483	5	34	241	93	372	1422

is randomly selected and its level is randomly changed to another permissible level. In a generic GA, factors mutate with an equivalent specified probability. Hence, the mutation rate does not incorporate other information gathered from prior knowledge about the system. For the SELC, we propose the use of prior information for generating mutation probabilities. For instance, suppose we know that the factor x_1 has a significant main effect and *no* significant two-factor interactions. Then, we will change the level of this factor to a new level, l, with probability p_l, where

$$p_l \propto \bar{y}(x_1 = l)$$

Next, suppose that factors x_1 and x_2 have a significant interaction. Then, the mutation should have a joint probability on x_1 and x_2. That is, the mutation will occur if either x_1 or x_2 is randomly selected. Factor x_1 will be set to level l_1 and factor x_2 to level l_2 with probability $q_{l_1 l_2}$, where

$$q_{l_1 l_2} \propto \bar{y}(x_1 = l_1, x_2 = l_2)$$

If the selected factor does not have significant main effects or interactions, then its value is changed to any admissible levels with equal probability. Note that, by the *effect sparsity principle*, only a few factors turn out to be important, and, by the *effect hierarchy* principle,[18] main effect and two-factor interactions are more important than others, which justifies the weighted mutation. A Bayesian variable selection strategy[19] is used to identify the significant effects.

Once a new run (or set of runs) is identified, it is created in the laboratory to evaluate the response function y, unless it is prohibited by the forbidden array. If a newly identified run is prohibited, then it is discarded, and another new offspring is generated.

ILLUSTRATION

To illustrate the SELC method, we have chosen a combinatorial library for which a number of compounds have already been created and percent inhibition values have been determined for an enzyme targeted in an antibacterial drug discovery program.

Overall, the combinatorial space has five possible substructures at position A, 34 at position B, and 241 at position C, spanning a total of 40 970 compounds (Figure 3). To explore this space, an initial combinatorial subset of 2114 compounds was created and screened. Using the results from this initial screen and scientific knowledge about the target, three subsequent combinatorial subsets were created and screened. In total, 2483 compounds (6% of the total combinatorial space) were created. Table 2 summarizes the new substructure space explored by each iteration and provides insight into the historical optimization process. In general, each iteration explored small rectangular subsets of the combinatorial space. Notice, in iteration 1 all five substructures were explored for position A, 26 of 34 substructures were explored for position B, and 164 of 241 substructures were explored for position C. However, only 9.9% of the compounds from the 5 × 26 × 164 space were created. In iteration 2, no new substructures were explored for position A (all were explored in iteration 1), but seven new substructures were explored for position B, and nine new substructures were explored for position C.

Table 3. Desired Compound Characteristics for Combinatorial Chemistry Example

reactive matched	<1
risky matched	<3
molecular weight	<500
rotatable bonds	<10
rule of 5	<2
aromatic ring count	<5
polar surface area	<140
cLogP	<5

Table 4. Number of Active Compounds by Each Historical Iteration

iteration	no. active (total)	no. of active with all desired characteristics
1	3 (2114)	2
2	26 (208)	6
3	11 (128)	5
4	0 (33)	0
overall	40 (2483)	13

For the response of interest, compounds are considered active if their percent inhibition values are greater than 40. In addition to finding active compounds, we seek to find compounds that fall within the constraints of Table 3. These constraints include the chemical properties: chemical reactivity, occurrence of toxicologically risky chemical features, molecular weight, number of rotatable bonds, violations of the Rule of 5,[20] aromatic rings, calculated polar surface area, and LogP (hydrophobicity).

Of the 2483 compounds that were created, 69% have all desired characteristics, while 31% have one or more undesirable characteristics. While the second iteration successfully finds 26 active compounds, only 6 of these have all of the desired characteristics (Table 4). And overall, the number of active compounds found that meet all desired characteristics is low (0.5%). Using the SELC method, we will attempt to identify substructures associated with active compounds and with the scientific intuition used to generate compounds in iterations 2, 3, and 4.

Implementing the SELC Method. As recommended in Mandal et al.,[15] we will initialize the experiment using an orthogonal array.[21] An orthogonal array of strength t, denoted by OA(N, $s_1^{m_1} s_2^{m_2} \ldots s_\gamma^{m_\gamma}$, t), is an $N \times m$ matrix, $m = m_1 + m_2 + \ldots + m_\gamma$, in which m_i columns have $s_i (\geq 2)$ levels such that, for any t columns, all possible combinations of levels appear equally often in the matrix. Usually when we refer to an array of strength 2, the index $t = 2$ is dropped for notational simplicity. An orthogonal array has two primary benefits. First, an OA is efficient because it requires fewer units (compounds) to precisely estimate factor (substructure) effects. An OA can also effectively estimate interactions among factors. From Table 2, iteration 1 explored 5

Table 5. Regression Analysis of Substructure Position for the Original Combinatorial Data and for the Orthogonal Array Subset

	estimate	SE	*t* value	Pr(>|*t*|)
Original Data				
(intercept)	1.10	0.31	3.50	0.000
A	−36.16	12.30	−2.94	0.003
B	−0.22	6.86	−0.03	0.975
C	69.87	16.31	4.29	0.000
A^2	−1264.76	413.31	−3.06	0.002
B^2	−387.17	296.81	−1.30	0.192
C^2	−1115.25	385.72	−2.89	0.004
AB	−638.24	315.01	−2.03	0.043
AC	−260.33	417.99	−0.62	0.533
BC	−170.85	352.54	−0.49	0.628
Orthogonal Array Subset				
(intercept)	1.58	0.65	2.41	0.017
A	−12.37	6.16	−2.01	0.046
B	4.06	5.94	0.68	0.495
C	37.88	13.22	2.87	0.005
A^2	−268.62	88.29	−3.04	0.003
B^2	42.40	93.70	0.45	0.651
C^2	−135.48	65.85	−2.06	0.041
AB	−43.05	78.49	−0.55	0.584
AC	78.24	75.94	1.03	0.304
BC	−85.42	76.15	−1.12	0.263

substructures for A, 26 for B, and 164 for C. Due to these constraints, we use a 625 run $25 \times 25 \times 25$ orthogonal array, spanning positions A, B, and C, respectively. Because position A has only 5 levels, we collapse the 25 levels at position A of the orthogonal array down to five levels through merging.[18] For positions B and C, we chose the 25 substructures with the highest frequency of occurrence. Of the 625 substructure combinations suggested by this OA, 241 compounds have been created in iteration 1. Moreover, only 173 of these meet the requirements in Table 3, and we will use this constrained subset as our initial information for the SELC algorithm. Although this subset contains information for only 28% of compounds from the original OA, these contain similar information about substructure importance as the entire compound set from iteration 1. Table 5 presents a regression analysis on both the entire set of compounds from iteration 1 and the subset by the OA. Clearly, positions A and C are identified as significant factors in explaining the response for both the iteration 1 data and the OA subset.

Upon conducting the initial analysis, we construct the forbidden array. Ultimately, we desire to find highly active compounds that meet the desired characteristics presented in Table 3. As a surrogate to a scientist's knowledge, we have used the historical data to identify undesirable combinations of characteristics. Hence, our forbidden array includes the worst compound and all compounds from the historical screen with two or more undesirable characteristics (n=70).

The forbidden array will also have order 2, which means that any pairwise interactions present in the forbidden array will not be allowed to be created in future experiments. For example, the worst compound in the initial experiment has substructure 3 in position A, substructure 10 in position B, and substructure 3 in position C. Hence, in future experiments we will not allow any compounds with substructure 3 in position A and 10 in position B, 3 in position A and 3 in position C, and 10 in position B and 3 in position C.

After selecting the forbidden array, we choose mutation probabilities for each substructure at each position. These probabilities are weighted according to the average substructure performance in iteration 1 (see Table 6). Each substructure receives the same baseline weighted mutation probability ($0.2 \times 1/4$). Then, for those substructures with a positive average response, an additional weight is added. For example, the additional weight added to substructure 2 is ($0.92/(0.92 + 0.8) = 0.53$). Substructures for positions B and C are treated similarly. Using the forbidden array and these weighted mutation probabilities, we now use the SELC method to suggest 200 new compounds to create and screen.

Table 6. Average Response and Weighted Mutation Probabilities for Each Substructure at Position A

substructure	av response	weighted mutation probability
1	−1.77	$= 0.2 \times 1/4 + 0 \times 3/4$
2	0.92	$= 0.2 \times 1/4 + 0.53 \times 3/4$
3	0.80	$= 0.2 \times 1/4 + 0.47 \times 3/4$
4	−1.30	$= 0.2 \times 1/4 + 0 \times 3/4$
5	NA	$= 0.2 \times 1/4 + 0 \times 3/4$

Table 7. Breakdown of SELC Suggested Compounds by the Original Iteration Number

iteration	no. active (total)	no. of active with all desired characteristics
1	0 (102)	0
2	2 (8)	2
3	0 (3)	0
overall	2 (113)	2

Of the 200 compounds suggested by the SELC, 113 (56.5%) were screened in the original experiment, which is a substantial enrichment over the percent that we would expect by random selection ($5.2\% = (2114/40970) \times 100$) (Table 2). Hence, the enrichment provided by the SELC is strong evidence that the method is identifying meaningful information. Of the 113 compounds identified by the SELC that had been screened, 102 were from iteration 1, 8 were from iteration 2, and 3 were from iteration 3 (Table 7). Additionally, 88 (77.9%) of these compounds had all of the desired characteristics. Moreover, both actives that were selected had all desired characteristics.

While the SELC identified only two active compounds, it has identified substructures associated with highly active compounds. For the 113 compounds that were screened, all 3 substructures of position A that are associated with the observed highly active compounds were identified, 11 of 12 substructures for position B were identified, and 6 of 16 substructures for position C were identified. These ratios were even better for the substructures associated with highly active compounds that met all of the desired characteristics. For these compounds, SELC identified all 3 substructures for position A, all 6 substructures for position B, and 5 of 10 substructures for position C. Because the SELC uses a weighted mutation scheme for each position, it will be able to efficiently create highly active compounds with desired characteristics.

In this example, we were able to use the SELC method to identify substructures associated with highly active compounds, identify highly active compounds, and improve upon the percent of compounds with desirable characteristics. Most importantly, this example illustrates that the SELC has the potential to save a significant amount of resources: only 15% of the original resources ($100 \times (173 + 200)/2483$) were used to find active compound matter and to enrich our knowledge of the chemical space.

CONCLUSIONS

Improvements in technologies have enabled scientists to create compounds more efficiently and screen compounds more rapidly, providing a wealth of information about a wide range of chemical space for which to make decisions. Despite the improvement in technology, constraints still exist for allowing scientists to fully utilize these technologies and the information generated by them. For example, a large combinatorial array can be designed, but resources are often limited to creating only a fraction of the desired space. Compounds can also be screened across a number of arrays, but the ability to fully understand complex underlying relationships is limited. Hence, there is an immediate need to couple these new technologies with statistical and optimization tools to enable scientists to harness the maximum amount of information and make informed decisions.

Both methods suggested in this work, desirability functions and SELC, can be quickly and easily implemented and coupled to existing technology to enhance decision making. Moreover, both methods can incorporate expert knowledge to allow the procedures to identify more relevant compounds. The desirability function, for instance, can be weighted to emphasize information from some screens, while de-emphasizing information from other screens. Likewise, the SELC method incorporates prior knowledge through the forbidden array, which is a unique characteristic of this optimization technique. In addition, the SELC method is guided by learned information through the weighted mutation scheme. Because modified desirability functions and the SELC incorporate expert knowledge to guide the optimization process, each can be easily influenced. For desirability functions, we suggest that the user select a scientifically defendable range of parameter values prior to calculating the desirabilities. In the case of SELC, we suggest that the user select only compounds that are scientifically defendable for placement into the forbidden array.

In both of the practical examples presented here, desirability functions and the SELC effectively identified important compounds and characteristics of important compounds. In the case of the SELC, significantly fewer resources were required to find potent chemical matter.

These tools combined with appropriate scientific knowledge have the potential to improve the efficiency of identifying effective and safe chemical entities with desirable ADME properties. Moreover, both methods can be used to enrich compounds libraries.

ACKNOWLEDGMENT

The research is supported by National Science Foundation grant DMS-0305996 (A.M. and J.W.) and the grant from the University of Georgia Research Foundation (A.M.).

Supporting Information Available: Data used for the desirability example, the initial design and response, the forbidden array, and the follow-up experiment for the combinatorial chemistry example. This material is available free of charge via the Internet at pubs.acs.org.

REFERENCES AND NOTES

(1) Cleaves, K. S. Automating R&D. *Mod. Drug Discovery* **2004**, *7* (6), 37−39.

(2) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: London, 2003.

(3) Gasteiger, J.; Engel, T. *Chemoinformatics: a Textbook*; Wiley-VCH: Weinheim, 2003.

(4) Rouhi, A. M. Custom Synthesis for Drug Discovery. *Chem. Eng. News* **2003**, *81* (7), 75−78.

(5) Myers, R. H.; Montgomery, D. C. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed.; John Wiley & Sons: New York, 2002.

(6) Del Castillo, E.; Montgomery, D. C. A Nonlinear Programming Solution to the Dual Response Problem. *J. Qual. Technol.* **1993**, *25*, 199−204.

(7) Khuri, A. I.; Conlon, M. Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions. *Technometrics* **1981**, *23*, 363−375.

(8) Pignatiello, J. J., Jr. Strategies for Robust Multi-response Quality Engineering. *IIE Trans.* **1993**, *25*, 5−15.

(9) Vining, G. G. A Compromise Approach to Multiple Optimization. *J. Qual. Technol.* **1998**, *30*, 309−313.

(10) Harrington, E. C., Jr. The Desirability Function. *Ind. Qual. Control.* **1965**, *21*, 494−498.

(11) Derringer, G.; Suich, R. Simultaneous Optimization of Several Response Variables. *J. Qual. Technol.* **1980**, *12*, 214−219.

(12) Del Castillo, E.; Montgomery, D. C.; McCarville, D. R. Modified Desirability Functions for Multiple Response Optimization. *J. Qual. Technol.* **1996**, *28*, 337−345.

(13) Kros, J. F.; Mastrangelo, C. M. Comparing Methods for Multi-Response Design Problem. *Qual. Reliab. Eng. Int.* **2001**, *17*, 323−331.

(14) Ortiz, F.; Simpson, J. R.; Pignatiello, J. J.; Heredia-Langner, A. A Genetic Algorithm Approach to Multiple-Response Optimization. *J. Qual. Technol.* **2004**, *36*, 432−450.

(15) Mandal, A.; Wu, C. F. J.; Johnson, K. SELC: Sequential Elimination of Level Combinations by Means of Modified Genetic Algorithms. *Technometrics* **2006**, *48*, 273−283.

(16) Holland, J. M. *Adaptation in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975.

(17) Holland, J. M. *Adaptation in Natural and Artificial Systems*; The MIT Press: Cambridge, MA, 1992.

(18) Wu, C. F. J.; Hamada, M. *Experiments: Planning, Analysis, and Parameter Design Optimization*; John Wiley & Sons: New York, 2000.

(19) Chipman, H.; Hamada, M.; Wu, C. F. J. A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing. *Technometrics* **1997**, *39*, 372−381.

(20) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(21) Hedayat, A. S.; Sloane, N. J. A.; Stufken, J. *Orthogonal Arrays: Theory and Applications*; Springer-Verlag: New York, 1999.

CI600556V