A Factorial Design To Optimize Cell-Based Drug Discovery Analysis

Bingming Yi,[†] Jacqueline M. Hughes-Oliver,*[,†] Lei Zhu,[‡] and S. Stanley Young[§]

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, and Cheminformatics and Statistics Unit, GlaxoSmithKline, Research Triangle Park, North Carolina 27709

Received March 6, 2002

Drug discovery is dependent on finding a very small number of biologically active or potent compounds among millions of compounds stored in chemical collections. Quantitative structure−activity relationships suggest that potency of a compound is highly related to that compound's chemical makeup or structure. To improve the efficiency of cell-based analysis methods for high throughput screening, where information of a compound's structure is used to predict potency, we consider a number of potentially influential factors in the cell-based approach. A fractional factorial design is implemented to evaluate the effects of these factors, and lift chart results show that the design scheme is able to find conditions that enhance hit rates.

INTRODUCTION

High throughput screening (HTS) technology is routinely used to identify lead molecules in the discovery of a new drug. During HTS, large collections of compounds are tested for potency with respect to one or more assays. In reality, only a very small fraction of the compounds in a collection will be potent enough to act as lead molecules in later drug discovery phases. Moreover, the expense of further investigation of leads places constraints on the number and quality of leads that will be pursued. The consequence is that out of a very large collection of compounds only a select handful will need to be identified for further study. Clearly, testing hundreds of thousands of compounds, one at a time, can be very wasteful, both in terms of time and money. Cost-effectiveness is critical for HTS programs.

For this reason, biologists, chemists, computer scientists, and statisticians have studied and searched for structure−activity relationships (SARs), which are built on the belief that chemical structure is highly related to potency of compounds. However, the models that relate biological potency and molecular structure are usually unclear and are often made more complex because molecular activity follows more than one mechanism. Nevertheless, it is well agreed that two molecules with fairly close molecular structure will have similar biological potency.[1] Based on this belief, many kinds of descriptors of molecular structure have been computed with the goal of improving HTS efficiency. Descriptors such as atom pairs, topological torsions, and fragments[2,3] have been useful in recursive partitioning analysis and/or pooling methods for HTS.[4−7] Other researchers use the continuous BCUT descriptors given by Pearlman and Smith[8] and derived from Burden.[9] Lam, Welch, and Young[10,11] propose a cell-based analysis (we call it the method LWY) using BCUT numbers.

The purpose of this paper is to improve the LWY methodology for screening large collections of compounds with respect to biological potency based on testing only a small fraction of the compounds. The approach develops a relational model between biological potency of a compound and its chemical structure. This relational model is built upon a relatively small selection of compounds (called the training set) for which both chemical structure and biological potency must be known or available. For the remainder of the compound collection (called the validation set), only information on chemical structure of compounds is needed. Compounds in the validation set will then be ranked according to predicted biological potency, using the relational model developed from the training set, and testing will proceed based on ranks until a desired number of potent compounds have been identified to serve as leads in the remainder of the drug discovery process.

Lam et al.[10,11] have already demonstrated that great gains can be achieved using method LWY. We argue that even greater gains are possible by tuning the method as well as incorporating modifications that make the method less computationally intensive and hence more feasible. We illustrate these methods using a data set from the National Cancer Institute (NCI). The second section contains a short description of the NCI data. The third section describes the cell-based method LWY in three subsections. The fourth section explores possible modifications to LWY using a fractional factorial experimental approach. The "best" cell-based methods are selected in the fifth section, and confirmatory results are shown in the sixth section. We close with a summary and discussion in the final section.

NCI DATA, AUGMENTED WITH BCUT DESCRIPTORS

The NCI maintains databases for the purpose of accelerating research on treatment of HIV/AIDS. One such database is located at http://dtp.nci.nih.gov/docs/aids/aids _data.html. It provides screening results and chemical structure on about 32 000 compounds for a specific antiviral assay. The chemical structure has been provided electronically in the form of a connection table giving the atoms and how they are bonded. These structures have been converted into quantitative features better suited for use in deriving quan-

* Corresponding author phone: (919)515-1954; fax: (919)515-1169; e-mail: hughesol@stat.ncsu.edu.
[†] North Carolina State University.
[‡] Cheminformatics, GlaxoSmithKline.
[§] Statistics Unit, GlaxoSmithKline.

10.1021/ci025509n CCC: $22.00
Published on Web 08/09/2002

titative structure−activity relationships (QSARs). In this paper we use BCUTs to numerically characterize the chemical structure.

BCUT numbers are eigenvalues from connectivity matrices derived from the molecular graph.[8,9] These numbers can be defined according to a variety of atomic properties such as size, atomic number, charge, etc. A different BCUT descriptor is created for each atomic property selected. More than 60 BCUT descriptors are in common use, but the high degree of multicollinearity existing between them has resulted in fairly common use of only six relatively uncorrelated BCUT descriptors. Given a particular atomic property of interest, a connectivity matrix from a molecular graph is constructed as a square matrix with dimension equal to the number of heavy (non-hydrogen) atoms. The property is placed along the diagonal for each heavy atom, while off-diagonal elements measure the degree of connectivity between two heavy atoms. Because eigenvalues are matrix invariant, they measure properties of the molecular graph. Being functions of all the heavy atoms in the molecule, the eigenvalues are thought to represent the properties of the molecule as a whole.

We were able to compute BCUTs for 29 812 compounds from the full NCI database, so our analysis is limited to this as the full NCI data set. Only 608 (2.04%) of the 29 812 compounds were potent for this antiviral assay. The histograms shown in Figure 1 indicate that BCUTs 1−6 capture different features as evidenced by the differences in central tendency and dispersion. There exist relatively low correlations among these six BCUTs, and the correlation coefficients range from −0.49 to 0.46, as illustrated in the following correlation matrix:

	BCUT1	BCUT2	BCUT3	BCUT4	BCUT5	BCUT6
BCUT1	1.00	−0.22	−0.42	−0.16	0.46	0.32
BCUT2		1.00	0.20	0.32	−0.25	−0.49
BCUT3			1.00	0.08	−0.25	−0.17
BCUT4				1.00	−0.03	−0.16
BCUT5					1.00	0.24
BCUT6						1.00

As previously mentioned, we need to select a training set from which the relational model will be developed. However, to assess issues of sensitivity of the method to the particular choice of training set, we select 20 pairs of training and validation sets. Each training set contains 4096 compounds (13.7% of the collection), so the associated validation set contains 25 716 compounds. These 20 training sets were selected by randomly sampling from the 29 812 compounds. They contain between 82 (2.00%) and 87 (2.12%) potent compounds.

Clearly, this is a case where biological potency is known for all compounds in the collection, so prediction is not needed. However, it provides an excellent test case for determining the effectiveness of a method. For the purpose of using the relational model developed from the training set to predict potency, activities are assumed unknown for all compounds in the validation set. After the model is built using the training set, the potencies of the compounds in the validation set are "revealed" and used to determine the quality of the relational model.

METHOD LWY, A CELL-BASED METHOD

Method LWY, as proposed by Lam et al.,[10,11] is based on dividing the six-dimensional BCUT space into cells. These cells are expected to be formed "good" enough to capture active regions that are dense with potent compounds. Once active regions have been identified, a sensible prediction approach is to test only compounds that fall in these active regions. This binning approach raises several questions, however. In the following subsections we describe the components of method LWY:

- How are cells formed? How do we divide a six-dimensional descriptor space into small enough regions to capture a high proportion of potent compounds without making the regions so small that they are too numerous to handle or contain too few compounds to be useful? This is discussed in the first subsection.
- How are cells classified as being good? Is a cell containing four potents among five compounds as good as a cell containing two potents among two compounds? This is discussed in the second subsection.
- The relational model developed in the two previous steps must now be applied to the validation set. How is this done? This is discussed in the third subsection.

Forming Cells. The six-dimensional BCUT space may be viewed more simply as low-dimensional subspaces, for example, all one-dimensional, two-dimensional, and three-dimensional subspaces. By focusing on these lower dimensions, LWY is able to create a very fine resolution of cells in order to find active regions. Moreover, these cells can be shifted to effectively multiply the number of cells available.

In the most extreme case, the six-dimensional space can be marginalized to create six one-dimensional (1-D) subspaces, one for each of the six BCUTs. Any compound falling in the six-dimensional space will also fall in exactly one cell of each of the six 1-D subspaces. The disadvantage of 1-D subspaces is that information on the relationship with other BCUTs is lost; the advantage is that we have a much smaller space that will accommodate a finer resolution of cells. All 1-D subspaces are divided into 64 disjoint bins. To diminish the impact of outliers (common in chemistry data sets), the first bin is formed by containing the lowest 1% of values for that BCUT, and the last bin is formed by containing the highest 1% of values. The remaining 62 bins have equal width between the extreme bins, irrespective of content.

To form two-dimensional (2-D) subspaces, each dimension has eight bins that are formed by amalgamating eight 1-D bins into a single bin. This maintains the total number of cells in the 2-D subspace to be 64, as it was for the 1-D subspace. Figure 2 illustrates this amalgamating process. It is easy to see that there are $\binom{6}{2} = 15$ 2-D subspaces.

Similarly, three-dimensional (3-D) subspaces are formed by amalgamating 16 1-D bins to create four bins in each dimension. There are $\binom{6}{3} = 20$ 3-D subspaces. Four-, five-, and six-dimensional (sub)spaces are excluded from use.

Beyond considering 1-D, 2-D, and 3-D subspaces, shifted cells are also created. Each dimension of a subspace is shifted to the right by half the length of one bin. So for each 1-D subspace, another 1-D subspace with 64 shifted bins is formed after shifting. For each 2-D subspace, with two dimensions allowed to shift, a total of four subspaces are

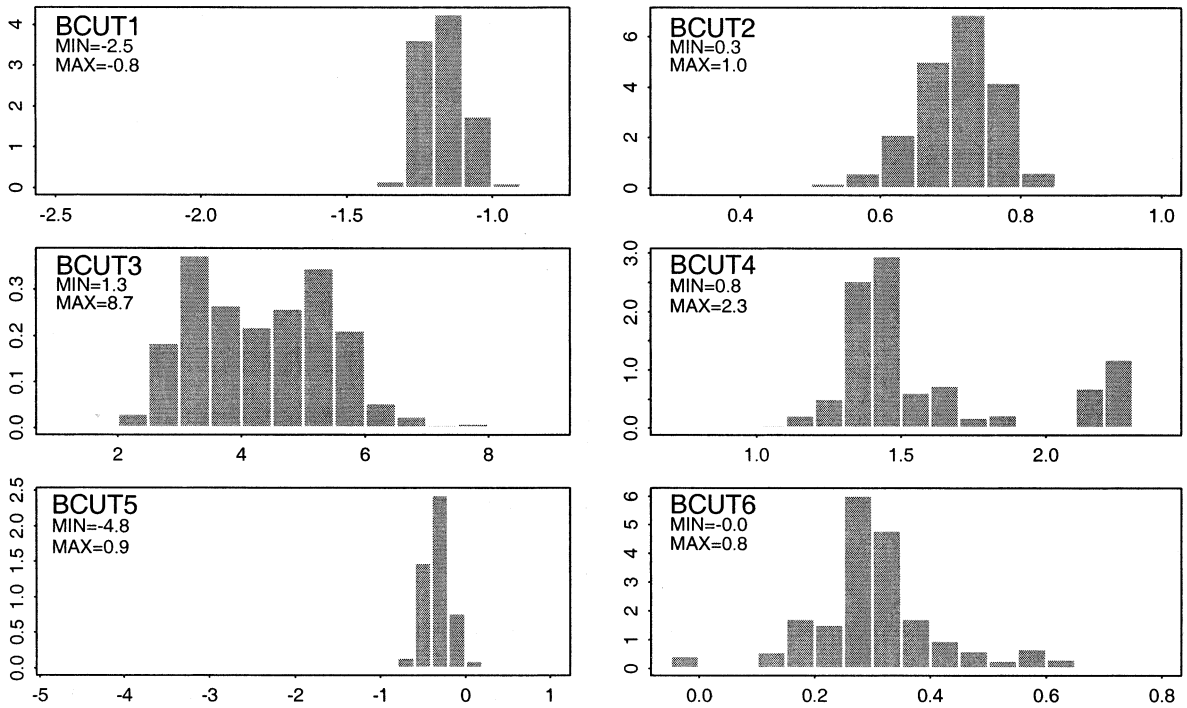

Figure 1. Histograms of six BCUT number descriptors.

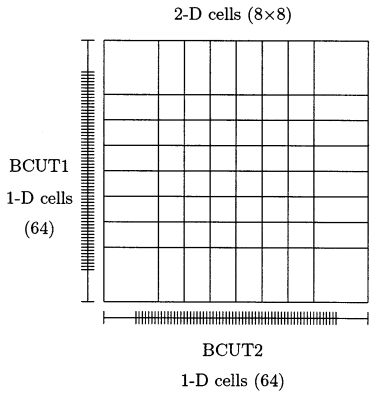

Figure 2. Forming 1-D and 2-D cells. There are 64 1-D cells where the extreme cells each contain 1% of the compounds and the middle 62 cells have equal width. There are also 64 2-D cells each of which is formed by collapsing 8 1-D cells from each dimension.

formed (including the original subspace without shifting). Similarly, for each 3-D subspace, a total of eight subspaces are formed after shifting.

Consequently, we have a total of $6 \times 2 + 15 \times 4 + 20 \times 8 = 232$ subspaces. Each subspace has 64 cells, so that we have a total of $232 \times 64 = 14\,848$ cells. It is important to remember that any given compound resides in every subspace. But in each subspace a compound resides in only one cell. Thus, we know each compound resides in 232 cells.

Identifying Good Cells. Method LWY applied to a training set from the NCI data yields 13 010 nonempty cells among the total of 14 848 cells. In these nonempty cells, about 43% contain at least one potent compound. Proportion quantiles of these "potent" cells are shown in Table 1. Clearly, a method is needed for identifying cells as being good or not. In method LWY, after all cells are formed, cells with two or more potent compounds will be separated from others and called the preliminary good cells. From now on, we will work on only these preliminary good cells.

Table 1. Quantiles for Potency Proportion of Cells Containing at Least One Potent Compound

quantile	min.	1%	5%	25%	50%	75%	95%	99%	max.
proportion	0.01	0.01	0.01	0.02	0.03	0.05	0.18	0.50	1.00

To further evaluate the preliminary good cells, a statistic is proposed to rank the cells. Now consider a cell with a total of n compounds where a of them are potent. If we assume the number of potent compounds follows the binomial distribution, we can find the 95% confidence limit for the potency proportion. (This proportion is called the hit rate when testing compounds in this cell: hit rate is the number of potent compounds identified divided by the number of compounds in the cell.)

The lower bound of the confidence interval can be used as a good criterion to know whether a cell is active. This lower 95% bound is labeled H_{L95} and defined as

$$H_{L95} = \min_p \{\Pr[Bin(n, p) \geq a] \geq 0.05\}$$

After the H_{L95} are computed for all preliminary good cells, we also need a cutoff value to separate good cells from the others. We use a permutation test to find the cutoff.

The training data contains 4096 compounds, each of which has a value of potency and six numerical BCUT descriptors. Suppose we randomly reorder the potencies of compounds but keep the BCUT numbers in their initial positions. This means that the potencies are randomly reassigned to the BCUT descriptors. Ideally, after this permutation no cells should be identified as good cells and large values of the H_{L95} of some cells are just the results of chance. The 95th-percentile of all such H_{L95} values obtained from permutations is labeled H_0. The value H_0 is used to distinguish between random variation and systematic variation. Returning to the true training set, all cells with H_{L95} larger than H_0 will be regarded as good cells.

It is computationally expensive to implement many permutations. Fortunately for the cell-based analysis, an equivalent way is to implement one permutation that reassigns the potencies in all 14 848 cells. This is much like doing thousands of random permutations or reordering the potency for each of 232 subspaces. Thus, we can take the 95th-percentile from one permutation as the cutoff H_0 to determine whether there are any real good cells in these preliminary good cells.

This is also a good way to determine if BCUT descriptors are relevant to activity. (Although widely used, BCUT numbers are rather abstract, and it is not clear if they contain structural information pertinent to biological activity.)

Prediction and Validation. Until now we have been working in the world of the training set. It is time to make predictions. Once good cells have been identified using the training set, we will select compounds residing in these good cells from the validation set. We will test these compounds one at a time. Which compounds should be tested first? This is very important since the order has considerable impact on hit rate results. For example, when we select 10 compounds to test sequentially, we have 10 chances to compute the hit rate after each test and obtain 10 hit rates. If all the 10 compounds selected happen to be inactive, the 10 hit rates are all zero. However, if the 10 compounds tested are all potent, the 10 hit rates are all 100%, which is the perfect result we desire.

So some criterion is needed to determine the order of testing. For this purpose, we will compute scores for the compounds in the validation set. We will determine the testing order by the scores. Frequency of occurrence of residing in good cells is a reasonable way of defining scores. For compound C_i, the score is defined as

$$S_i = \sum_{cell_k \text{ is good}} I(C_i \in cell_k)$$

where $I(A)$ is the indicator function that takes value one when A is true and zero otherwise.

Alternatively, we can apply a weight to each good cell and compute a new score for the compounds

$$S_i = \sum_{cell_k \text{ is good}} I(C_i \in cell_k) H_{L95,k}$$

where $H_{L95,k}$ is the H_{L95} value for $cell_k$.

For both scoring functions suggested above, we expect that the higher the score the greater the chance the compound will be potent, and thus it should have a higher priority to be among the first tested. So, we can rank the compounds in the validation set by the order of high score to low. These compounds will be tested by this order, and the hit rate results can be computed. Method LWY uses the latter as scores.

OTHER CELL-BASED METHODS

In the cell-based method proposed by Lam et al.,[10,11] many tuning factors may be investigated. For example, we can consider only 3-D subspaces rather than all 1-D, 2-D, and 3-D subspaces. If 3-D subspaces are enough to produce comparable hit rate results, then the computational burden could be reduced to consider only 160 subspaces instead of 232. So subspace is a factor to be considered and we examine two levels: 3-D subspaces only and all 1-D, 2-D, and 3-D subspaces. In fact, we decide on seven factors that could be further investigated to optimize LWY's cell-based method. These are summarized in Table 2 and fully discussed below.

Table 2. Factors for Creating Alternative Cell-Based Methods

factor	low	center	high	description			
A	−1	0	1	cell	64	216	729
B	−1	0	1	amalgamating	No1	No2	yes
C	−1		1	subspaces	3-D		1−, 2−, 3−D
D	−1		1	permutation	1		5
E	−1		1	preselection	no		yes
F	−1		1	percentile	95%		100%
G	−1	0	1	weights	$H_{L95}/\sqrt{n}$	H_{L95}	$\sqrt{H_{L95}}$

Factor A considers the number of cells into which a subspace is divided. Method LWY uses 64 cells always, for 1-D, 2-D, and 3-D subspaces. We consider two higher numbers of cells, 216 and 729. Is it beneficial to have a finer resolution of cells and is there a point beyond which no gains are realized? For 729 cells, 1-D subspaces are divided into 729 bins, 2-D subspaces are divided into 27 bins on each dimension ($27^2 = 729$), and 3-D subspaces are divided into 9 bins on each dimension ($9^3 = 729$). For 216 cells, things do not work out as cleanly. Notice that 216 cells can be obtained for 1-D and 3-D subspaces ($6^3 = 216$) but not for 2-D subspaces. We make a slight modification and use $15^2 = 225$ for 2-D subspaces. So factor A, called cell, has three levels: low (coded as −1), represents 64 cells in each subspace; center (coded as 0), represents 216 cells in each subspace; and high (coded as 1), represents 729 cells in each subspace.

Factor B, called amalgamating, considers whether amalgamating will be used in creating 2-D and 3-D cells from 1-D cells; see the subsection on forming cells. Factor B has three levels, where the high level (coded as 1) means amalgamating will be performed, as in method LWY. The other levels, low (coded −1) and center (coded 0), do not use amalgamating. For the low level, denoted "amalgamating = No1," the first and last bins in each dimension contain the maximum of 1% of all compounds or the inverse of the number of bins in that dimension. In other words, the number of compounds in the extreme bins is max(1%,(# of bins in that dimension)$^{-1} \times 100\%$). For example, for 3-D subspaces with 64 cells, each dimension has four bins, and the first and fourth bins will be formed to contain 25% of the compounds in the training set. But for 2-D subspaces with 64 cells, the first and eighth bins will be formed to contain 12.5% of compounds in the training set. For the center level, denoted "amalgamating = No2," the first and last bins in each dimension contain 1% of all compounds. For each dimension of 3-D subspaces with 64 cells, the second and third bin together will contain 98 of all compounds. It seems to gather too much data in the central cells and does not seem to be a very sensible approach, but we nevertheless compare it to the other levels.

Factor C considers whether all of the 1-D, 2-D, and 3-D subspaces will be used, as in method LWY, or only the 3-D subspace. The latter level is considered low (coded −1), and the former is considered high (coded 1) in this two-level factor.

Factor D has two levels to represent the number of permutation runs used to determine the cut point for good cells: low (coded as −1) represents one permutation and high (coded as 1) represents five permutations. As discussed in subsection Identifying Good Cells, LWY uses one permutation with the justification that the large number of cells makes it reasonable. We expect no difference between the two levels.

Factor E has two levels to indicate whether preliminary selection of good cells is performed. As discussed in subsection Identifying Good Cells, preliminary good cells with two or more potent compounds are first identified, and the really good cells are selected from among these cells. We doubt that preliminary selection is particularly effective. Low level (coded as −1) is no preliminary selection, and the high level (coded as 1) indicates preliminary selection is performed.

Factor F has two levels concerning what to use as the cutoff from the permutation test. Method LWY requires that in order to find a cutoff value to separate good cells from the others, we need to implement the permutation process and find the 95th-percentile H_{L95} value. We propose to increase the cutoff value by replacing the 95th-percentile with the maximum value after permutation. This will lead to a more strict selection for good cells. Low level (coded as −1) indicates use of the 95th-percentile, and the high level (coded as 1) indicates use of the maximum of 100th-percentile.

Factor G concerns the weight used in scoring compounds for predicted potency. In method LWY, weight H_{L95} is used to compute scores for compounds in the validation set. We propose using either $\sqrt{H_{L95}}$ or $H_{L95}/\sqrt{n}$ to replace H_{L95}. Low level (coded as −1) uses $H_{L95}/\sqrt{n}$, center level (coded as 0) uses H_{L95}, and high level (coded as 1) uses $\sqrt{H_{L95}}$.

We are not willing to consider all $3^3 2^4 = 432$ combinations of alternative cell-based methods, but we very much want to know what factors might be important in obtaining an effective method. To evaluate the main effects of these factors, we use the design procedure ADX in SAS to find an optimal main effects design. It gives the fractional factorial design with 21 design points shown in Table 3. Method LWY is not in this set of 21 but can be expressed and added as $A = -1$, $B = 1$, $C = 1$, $D = -1$, $E = 1$, $F = -1$, $G = 0$.

The resulting 22 cell-based methods (including method LWY) were applied to 20 training sets. Lift charts were obtained and analyzed using standard analysis of variance procedures.

Table 3. Optimal Design Points for Creating Alternative Cell-Based Methods

design	A	B	C	D	E	F	G
1	1	1	−1	1	1	−1	1
2	1	1	−1	−1	1	1	1
3	1	1	1	−1	−1	−1	0
4	1	0	1	1	1	1	0
5	1	0	−1	1	−1	1	0
6	1	0	1	1	1	−1	−1
7	1	−1	−1	−1	−1	−1	−1
8	0	1	1	−1	−1	−1	0
9	0	1	1	1	1	1	−1
10	0	0	−1	−1	1	−1	1
11	0	0	−1	−1	−1	1	−1
12	0	−1	1	1	−1	1	1
13	0	−1	−1	1	1	1	0
14	−1	1	−1	1	−1	−1	0
15	−1	1	−1	1	−1	1	−1
16	−1	1	−1	−1	1	1	−1
17	−1	0	−1	1	−1	−1	1
18	−1	0	1	−1	1	1	0
19	−1	−1	1	−1	−1	1	1
20	−1	−1	−1	−1	1	−1	0
21	−1	−1	1	1	1	−1	−1
LWY	−1	1	1	−1	1	−1	0

SELECTING THE "BEST" CELL-BASED METHOD

Using only 22 observed cell-based methods, our goal is to determine which of the 432 possible cell-based methods will be most effective for our assay-library combination. Realizing that no method will be uniformly optimal for all assay-library combinations, we propose the following: (a) use all 22 cell-based methods to screen the training sets; (b) assess the effectiveness of each screening method on each training set; (c) use analysis of variance to determine the best cell-based method for the training sets; (d) if the best cell-based method has not previously been observed, then apply it to the training sets to obtain confirmation of its effectiveness and to prepare for application to the validation sets; and then (e) apply this best method to the validation set.

Our assessment of the effectiveness of screening is done by the lift chart. Given input data, cell-based methods create relational models that output a ranked ordering of compounds, where ranking is according to predicted potency. Testing of compounds is done according to this ordering until "enough" potents have been identified. The relational model developed from the cell-based method, if successful, will assign high ranks to potent compounds, thus making it likely that the cumulative percent of potent compounds actually found, relative to the number of compounds tested, will be very high in early testing and will decrease to approach the average potency in the full data set. A lift chart plots the cumulative percent (hit rate) of potents found, relative to the number of compounds tested, as a function of the number of compounds tested. Obviously, the higher the hit rate, the better. Lift charts are directly related to accumulation curves[12] and cumulative recalls,[13] as discussed below, but are more amenable to the types of statistical analyses conducted here. Specifically, assumptions of normality and homogeneity of variances are very reasonable for hit rates.

Kearsley et al.[12] propose measures based on the accumulation curve. The accumulation curve plots the total number of potent compounds found ($A@n$) versus the total number of compounds tested (n). For example, $A@20$ is the number of potents found from testing the first 20 compounds. The lift chart at 20 tested compounds is simply $h_{20} = (A@20/20) \times 100\%$. In other words, the lift chart plots $((A@n/n) \times 100)\%$ versus n. The ideal case for the accumulation curve is the line "$A@n = n$," which corresponds to a flat line "$h_n = 100\%$" in the lift chart. The cumulative recall[13] plots the cumulative percent of potents found, relative to the total number of potents in the entire set, versus the total number of compounds tested. In other words, the cumulative recall plots $((A@n/C) \times 100\%)$ versus n, where C is a constant that represents the total number of potent compounds in the entire set. We see that the three types of plots are actually equivalent.

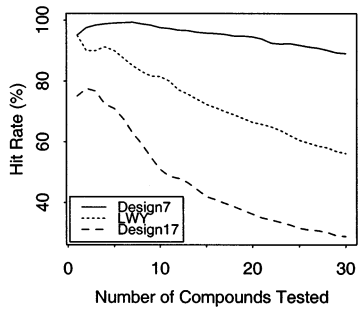

Figure 3. Lift charts for three methods, averaged over 20 training data sets.

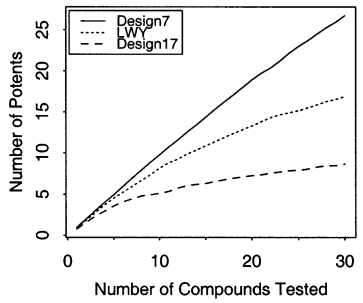

Figure 4. Accumulation curves for three methods, averaged over 20 training data sets.

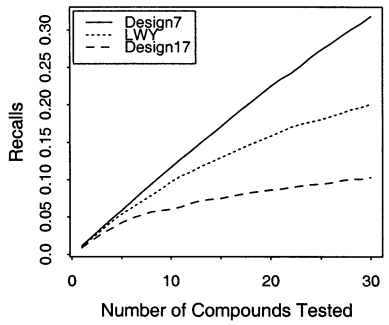

Figure 5. Cumulative recalls for three methods, averaged over 20 training data sets.

Figure 3 summarizes lift chart results from three cell-based methods, namely, LWY, design 7, and design 17. For each method and each training set, a lift chart is obtained. Because 20 training sets (that is, 20 replicates) are used, we average the 20 lift charts, and these are the plotted curves in Figure 3. It appears that design 7 is considerably better than LWY, while design 17 is considerably worse. Figures 4 and 5 show the same results as Figure 3, but in the forms of accumulation curve and cumulative recall, respectively. Of course, our designed experiment allows us to test for differences in light of uncertainty. But we must first determine what is the "response" of interest.

Kearsley et al.[12] consider a global enhancement based on *A*50, the number of compounds that must be tested until half the potents are found. Initial enhancement, another proposal based on *A@N* from a large database consisting of *M* compounds ($N = M/100$), is argued to be a more appropriate measure than *A*50 when only a small percent of a large database will be tested.

In obtaining our lift charts, we are interested in testing only a small percent of our collection of compounds. Because the training sets contain only 4096 compounds, we consider h_{10}, h_{20}, and h_{30} (equivalent to *A*@10, *A*@20, and *A*@30) as measures of assessment. Analysis of variance on the three responses h_{10}, h_{20}, and h_{30} show slightly different results, as summarized in Table 4. Possible correlation between responses is not addressed.

Table 4. Summary of ANOVA Results for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set[a]

	h_{10}		h_{20}		h_{30}	
source	F	*p*-value	F	*p*-value	F	*p*-value
model effect	37.3	≤0.0001	47.7	≤0.0001	62.6	≤0.0001
main effect A	141.1	≤0.0001	178.2	≤0.0001	229.1	≤0.0001
main effect B	26.9	≤0.0001	22.8	≤0.0001	33.1	≤0.0001
main effect C	36.0	≤0.0001	22.6	≤0.0001	21.2	≤0.0001
main effect D	0.5	0.4981	3.1	0.0767	14.0	0.0002
main effect E	0.4	0.5060	0.5	0.4757	0.0	0.9668
main effect F	2.4	0.1215	35.0	≤0.0001	65.5	≤0.0001
main effect G	11.0	≤0.0001	19.6	≤0.0001	18.7	≤0.0001

[a] F is the value of F statistics.

Table 5. Comparisons between Design 7, LWY, and Design 17 for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set[a]

	h_{10}	h_{20}	h_{30}
difference between design 7 and LWY	16.00	28.00	32.83
difference between LWY and design 17	30.50	30.25	27.34
LSD min. significant difference	7.58	7.59	6.53
HSD min. significant difference	13.95	13.97	12.03
Bonferroni min. significant difference	14.39	14.41	12.41

[a] Experimentwise levels of significance are all set at 5%.

Table 6. Mean Hit Rates, at All Levels of All Factors for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, 30 Compounds in the Training Set

	h_{10}			h_{20}			h_{30}		
factor	−1	0	1	−1	0	1	−1	0	1
A	70.4	86.3	94.1	53.8	67.7	81.0	45.0	56.9	70.9
B	89.0	77.3	84.4	72.5	61.6	68.5	63.1	51.9	57.8
C	79.9		87.3	64.6		70.5	55.2		60.0
D	83.2		84.0	68.6		66.4	59.6		55.6
E	83.2		84.0	67.1		68.0	57.6		57.6
F	84.5		82.6	71.2		63.9	61.8		53.4
G	86.9	84.4	79.5	72.8	67.3	62.5	62.1	57.3	53.4

Factor E is not important in creating cell-based methods for these responses and training sets. Especially when testing fewer compounds, factors D and F are not as important as factors A, B, C, and G. Multiple comparison testing suggests that hit rates for design 7 are significantly higher than for LWY, which in turn are significantly higher than for design 17 (see Table 5). Of the 22 cell-based methods listed in Table 3, the best method is design 7, which is profiled with 729 cells, amalgamating No1, using 3-D subspaces only, a single permutation, no preliminary selection, permutation percentile 95%, and weight $H_{L95}/\sqrt{n}$. The worst method is design 17, which has 64 cells, amalgamating No2, uses 3-D subspaces only, five permutations, no preliminary selection, permutation percentile 95%, and weight $\sqrt{H_{L95}}$.

Moreover, based partially on details presented in Table 6, it appears that level 1 of factor A (729 cells), level −1 of factor B (amalgamating No1), level 1 of factor C (1-, 2-, 3-D subspaces), level −1 of factor F (permutation percentile 95%), and level −1 of factor G (weight $H_{L95}/\sqrt{n}$) are

Table 7. Four Best Cell-Based Methods Predicted from the Statistical Analyses

factors							95% confidence intervals		
A	B	C	D	E	F	G	h_{10}	h_{20}	h_{30}
1	−1	1	−1	1	−1	−1	[103.2, 111.8]	[95.1, 103.6]	[85.9, 93.1]
1	−1	1	−1	−1	−1	−1	[102.2, 111.1]	[94.1, 102.9]	[85.9, 93.2]
1	−1	1	1	1	−1	−1	[104.2, 112.5]	[93.1, 101.3]	[82.1, 89.0]
1	−1	1	1	−1	−1	−1	[103.2, 111.8]	[92.1, 100.5]	[82.1, 89.1]
design 7							[94.9, 103.6]	[88.3, 97.0]	[81.1, 88.3]

Table 8. Global and Initial Enhancement for Methods LWY and OPT

method	*A*@250	IE
design 7	90	17.7
OPT	90	17.6
LWY	68	13.3

associated with better methods. In fact, four cell-based methods, as predicted best from the statistical analyses, are given in Table 7. The best among these four designs is denoted OPT and coded as (A,B,C,D,E,F,G) = (1, −1, 1, −1, 1, −1, −1). All four methods share the key features listed above, and they are not significantly different at the $\alpha = 0.05$ level. In fact, they differ only in levels for the mostly unimportant factors D and E. The ignorable impact of factors D and E is evidenced by the largely overlapping 95 confidence intervals. None of these four methods are included in the set of 22 methods listed in Table 3, and any of them could be selected as the best cell-based method. However, with one small exception, 95% confidence intervals for these fours methods also overlap the 95% confidence interval for design 7 (see Table 7), which means that the four methods are not significantly better than design 7. In our case, design 7 is selected to make prediction as the best cell-based method. In the circumstances when the methods predicted best are significantly better than the observed best, they should be chosen to make prediction.

CONFIRMATORY RESULTS

The design of experiments approach to selecting a best cell-based method for this assay-library combination allows us to conclude that design 7 can be used to give near-optimum performance. Applying design 7 to the 20 training sets, the observed mean responses are 97.5%, 94.5%, and 89.0% for h_{10}, h_{20}, and h_{30}, respectively. We are now ready to investigate the enhancements offered by this cell-based method compared to random testing and method LWY. To justify our claim that design 7 and OPT are comparable, we also provide details for method OPT. Enhancements are measured by applying the methods to the 20 large validation sets, each consisting of 25 716 compounds.

Following the approach of Kearsley et al.,[12] we use the measure of initial enhancement. Initial enhancement is the ratio of the actual *A*@250 for the method over the *A*@250 expected for random testing (250 × 520/25, 716 in our case), that is, $IE = A@250 \cdot (25{,}716/250 \times 520)$. These numbers are displayed in Table 8. Again, the benefits of design 7 and OPT are clear, with much better performance relative to both random testing and method LWY.

Figure 6 summarizes lift chart results for LWY, design 7 and OPT, averaged over the 20 validation sets. It also shows the expected lift chart from random testing. The message is quite clear, both design 7 and OPT offer substantial improvements and they make comparable predictions.

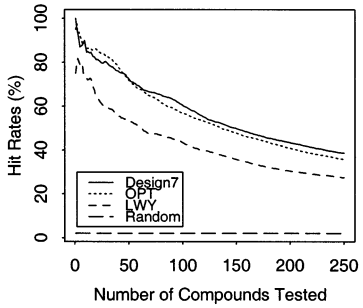

Figure 6. Lift charts for LWY, design 7 and OPT, averaged over 20 validation data sets.

SUMMARY AND DISCUSSION

Cell-based methodology is successful at improving HTS efficiency. It is much more efficient than random testing, and it has also been asserted to be more favorable than recursive partitioning.[10,11] This cell-based approach combines the power of multiple good cells with the incorporation of molecular structural descriptors. The novel point of this method is the investigation of SARs by implicitly accounting for the effects of descriptors; no explicit functional forms are suggested. These results reinforce the idea that compounds with similar structure will have comparable potency.

Gains in efficiency of cell-based methods compared to random testing are, however, sensitive to many factors of the process. It has been useful to implement an appropriate design of experiments for investigating the effects of these factors. From the results discussed above, we can see that careful selection of these factors can dramatically improve the screening efficiency. According to these results, the best method among the 22 methods in the main-effects design (including LWY) comes from design 7.

In design 7, only 3-D subspaces are required. This can lead to savings in computational time during the practical screening process. The gain can be quite substantial if the number of descriptors is increased. In this paper, only six of the more than 60 BCUT descriptors are used, and so the savings by considering only 3-D subspaces is that 72 fewer subspaces are needed; see the third and fourth sections. On the other hand, if 10 BCUT descriptors are used to develop the relational model, then 200 fewer subspaces are needed if only 3-D subspaces are used.

These benefits can, and should, be investigated for a variety of applications. We make no claims that design 7 will be best for all assay-library combinations. In fact, we strongly believe that the best cell-based method for one application may be horribly inefficient for another application and so it is imperative that the selection process be

application-specific. Indeed, this is what we consider as our major contribution: the ability to determine, for each assay-library combination, the best cell-based method for screening the bulk of the library. The best method is entirely determined from a relatively small training set, and no additional testing is required to assess many possible methods; the only additional cost is computing time.

Random selection of 20 pairs of training and validation sets offers realization of extreme behaviors, both positive and negative. For example, the lowest initial enhancement of design 7 over random testing is 16.3 and comes from training-validation pair 12. On the other hand, the highest initial enhancement of Design 7 over random testing is 23.5 and comes from training-validation pair 4. Enhancements quoted in Table 8 are averages over all 20 training-validation pairs, just as the enhancements quoted on page 124 of Kearsley et al.[12] are averages over 10 different probes within the same library. While our 20 training-validation pairs are not exactly equivalent to considering 20 separate assays, they do provide some information on the variability of the technique across applications.

There are, of course, areas in which our suggestions can be further investigated and possibly improved. It would be useful to have some sense of the relative importance of the BCUT numbers in predicting potency, but our current work makes no such attempts, with good reason. Cell-based methods isolate small regions in 1-, 2-, or 3-D subspaces where potent compounds reside. Examination of the BCUTs that index the important dimensions tells us which BCUTs are important, and the cell coordinates tell us the important ranges of those BCUT numbers. Examination of the potent compounds in these good cells and the nature of the particular BCUTs can point to specific atoms and substructures. Our experience is that this is complex, and the simple examination of the compounds generally illuminates the class of potent compounds in the subregion. The nature of the BCUTs provides information on atom-centric properties useful for modeling molecular characteristics. Having said all of this, the analysis here is largely aimed at getting high hit rates and interpretation of the molecular features is somewhat secondary. We should note that use of BCUTs for interpretable QSAR is considered problematic,[14] although recent work[15] is more positive.

Lam et al.[10,11,16] spend considerable effort in examination of various ways to evaluate the importance of a cell. Two things are important, the level of potency or hit rate, and some measure of the reliability of the estimated hit rate. There are obviously many ways to weight these factors. One of the ways recommended by Lam et al.[10,11,16] is to use a statistical lower bound, H_{L95}, on the proportion of potent compounds in the cell. Our work also uses H_{L95} in scoring cells, but others may find a different metric to be more desirable. Whatever the measure adopted, the cells can be ranked, and unscreened compounds can be tested either in the order of their occurrence or a predetermined number of compounds can be used to determine how many of the good cells will be tested.

Cell-based methods can be used as the statistical method for selecting compounds in a sequential screening scheme.[17] In sequential screening, an initial set of compounds is screened, and the results are used to determine a predictive model. A cell-based method could be used to make these predictions, and eventually a block of compounds are selected and tested. So either a preselected number of compounds would be assayed or all the compounds that fell into good cells. If the logistics of compound handling and screening are very good and if the assay is very expensive, it would make sense to screen very small, incremental sets of compounds and stop when the screening objectives are met, namely, when either a fixed number of potent compounds or compound classes are identified. This sequential screening paradigm might be even more effective if the scoring statistic H_{L95} is replaced by a more dynamic index.

In this work, we consider seven factors and specific levels for each in order to determine the best cell-based method. Actually, factors could be expanded to include additional levels or even to introduce new factors. An additional factor could be used to investigate the use of equal-width bins versus equal-frequency bins versus hybrid (a combination of equal-width and equal-frequency) binning. Lam et al.[10,11,16] report that they conducted extensive investigations of the effects of these various methods of binning and concluded that the hybrid approach was most effective. This finding could, however, be assay dependent, and it may be a good idea to include the factor in the analysis. Also, one could consider adding more levels for some factors. For example, 4-D and 5-D subspaces could be included as levels of factor C and factor G might also include other choices for the weights.

Selection of a main-effects design is somewhat limiting because two-factor interactions are only estimable when some main effects are negligible. If one is willing to entertain more runs, that is, observe more cell-based methods, larger designs that accommodate estimation of interaction effects could also be pursued. For example, assuming that two of the factors in the NCI data application were negligible (which is true of factors E and D), a design that allows estimation of the five remaining main effects and all 10 two-factor interactions contains 44 design points.

The design of experiments approach presented here is general enough to be applicable to many assay-library combinations, yet specific enough to result in significant improvements for a particular assay-library application.

ACKNOWLEDGMENT

This work was supported in part by a grant from the National Science Foundation, award number DMS-0072809. We also gratefully acknowledge Lap-Hing Raymond Lam at GlaxoSmithKline for providing the NCI data, early access to his dissertation, and other help.

REFERENCES AND NOTES

(1) McFarland, J. W.; Gans, D. J. On the Significance of Clusters in the Graphical Display of Structure−Activity Data. *J. Med. Chem.* **1986**, *29*, 505−514.

(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan R. Atom Pairs as Molecular Features in Structure−activity studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(3) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(4) Hawkins, D. M.; Young, S. S.; Rusinko A. Analysis of a large structure−activity data set using recursive partitioning. *Quantitative Structure−Activity Relationship* **1997**, *16*, 296−302.

(5) Young, S. S.; Hawkins, D. M. Using Recursive Partitioning to Analyze a Large SAR Data Set. *Structure−Activity Relationship and Quan. Structure−Activity Relationship* **1998**, *8*, 183−193.
(6) Rusinko A.; Farmen M. W.; Lambert C. G.; Brown P. L.; Young S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *38*, 1017−1026.
(7) Zhu, L.; Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **2001**, *57*, 922−930.
(8) Pearlman R. S.; Smith K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(9) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(10) Lam, R. L. H. Design and Analysis of Large Chemical Databases for Drug Discovery; Ph.D. Dissertation, University of Waterloo, 2001.
(11) Lam, R. L. H.; Welch W. J.; Young, S. S. Cell-Based Analysis for Large Chemical Databases. *Technometrics* **2002**, submitted.
(12) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(13) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Molecular Graphics . Modelling* **2000**, *18*, 343−357.
(14) Pearlman R. S.; Smith K. M. Novel Software tools for chemical diversity. *Perspectives Drug Discovery Design* **1998**, *9−11*, 339−353.
(15) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(1), 11−20.
(16) Lam, R. L. H.; Welch W. J.; Young, S. S. Uniform Coverage Designs for Molecule Selection. *Technometrics* **2002**, *44*, 99−109.
(17) Engels M. F.; Venkatarangan P. Smart screening: Approaches to efficient HTS. *Current Opinion Drug Discovery &Development* **2001**, *4(3)*, 275−283.

CI025509N