Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery†

S. Stanley Young,‡,§ Charles F. Sheffield,‡,⊥ and Mark Farmen*,‖

Glaxo Wellcome Inc., Five Moore Drive, Research Triangle Park, North Carolina 27709, and Department of Statistics, The Ohio State University, Columbus, Ohio 43210

Received March 21, 1997⊗

Pharmaceutical companies have a large inventory of compounds for screening projects. This inventory is valuable. The optimal use of inventory requires predictions of the largest potencies that will be observed in a large potential screening set. These predictions can be made using potency values observed on a large, but smaller set of compounds using statistical methods for analyzing extreme values. The break-even point between screening and synthetic modification of lead molecules is assessed and shown to depend critically on the cost of screening and synthetic modification. The validity of the extreme value theory approach is assessed using a large set of single point assay values.

INTRODUCTION

Inability to predict extreme outcomes can lead to wasted time and money. In Holland, engineers need to predict the highest floods likely to occur in the future so that their dikes will not be overtopped. However, if they are built too high, then money is wasted. Insurance companies need to predict the most costly settlements expected to occur over the policy lifetime since costs are largely determined by the largest claims. In addition, the costs must not be overestimated because setting premiums at too high of a level will result in loss of customers to competitors.

Drug discovery proceeds in two stages.[1] First, a lead molecule is found which is modestly potent. Second, this lead molecule is optimized by synthetically removing and adding parts. Typically, the initial lead is found by screening large collections of compounds from inventory. Alternatively, combinatorial synthesis[2−5] has come into recent use to create large collections of compounds to screen, chemical libraries. In either case, an important question is how many compounds should be screened from inventory, or how many library compounds should be synthesized before optimization is commenced by molecular modification. By examining the potency of a random sample of compounds, the expected magnitude of several of the most potent compounds in the entire collection (actual or virtual) can be estimated. Then, various strategies can be devised by taking into account the costs and expected gains from screening verses molecular modification so that a cost effective strategy can be devised. Why not just screen everything? At some point, it will be more cost effective to stop screening and start molecular modification. It would be much better to discover two drugs for the same price as one. In addition, it is important that an inventory of compounds is not being used inefficiently to the detriment of future drug discovery.

First, cost estimates of a compound collection are developed. A pharmaceutical company's inventory can be worth on the order of 100 million dollars, implying that mismanagement can be quite costly. In addition, the costs of constructing a one million compound library when a one hundred thousand compound library is adequate can also be assessed. Statistical extreme value theory can be used to estimate the maximum potency of compounds in a collection. Although mathematically intricate, new personal computer software makes this technology readily available. The methods are illustrated using real data. Next, a discussion is given of the importance of our results for cost effective drug discovery. Finally, a large collection of single point assay values is used to validate the extreme value theory approach for potency prediction.

COST OF A COMPOUND COLLECTION

It is essential to have a collection of diverse compounds to screen when an assay for a new biological target is developed.[6] For the long-term viability of the enterprise, it is also important that compound supplies not be wasted so that future screening activity is not deprived. Compounds are kept in inventory and used when the need arises. Such inventory compounds are dispensed to three areas: (1) specific drug projects; (2) Low- (moderate-) throughput screens; (3) High throughput screens.

The important distinction is that compounds going to specific drug projects are specifically made in an attempt to optimize the properties of a lead molecule, whereas compounds from inventory are sent to low- and high-throughput screens in an attempt to find new lead molecules.

Small organic molecules are acquired by companies from four sources: (1) directed synthesis for particular projects; (2) commercial sources such as the Fine Chemical Directory; (3) universities; (4) combinatorial synthesis. The cost varies considerably with the source. It stands to reason that the project should bear most of the cost for compounds specifically synthesized for it. Compounds from the other three sources should be cost effective with respect to the diversity they add to the current inventory. A company should not pay much for a compound which is very similar to one

† This paper is based on material presented at the First International Conference on Advanced Pharmaceutical Substance Screening sponsored by The Austrian Society for Animal Cell Culture, 1993.

‡ Glaxo Wellcome Inc.

§ E-mail address: young~ss@glaxowellcome.com.

⊥ E-mail address: sheffield~cf@glaxowellcome.com.

‖ The Ohio State University. E-mail address: farmen@stat.mps.ohio-state.edu.

⊗ Abstract published in *Advance ACS Abstracts,* September 1, 1997.

S0095-2338(97)00224-2 CCC: $14.00

Table 1. Cost Allocation in Dollars for a Typical Compound

compound use	labor	overhead	materials
primary screen	800	3000	20
HTS	200	750	80
inventory	200	750	200

already in its inventory, and it makes sense to pay a premium for a compound that is dissimilar.

An inventory of compounds for screening should necessarily be large and structurally diverse. A large company might have a tremendous inventory of compounds, but it is likely that many were synthesized for specific drug projects and constitute groups of compounds that are very similar to one another. There may be many steroids, β-lactams, etc., and hence the number of substantially diverse compounds in a collection is usually much less than the size of the collection. The diversity of compounds in a collection can be characterized using similarity measures and clustering,[7−9] but those questions will not be covered here.

Assuming that nothing is free, the cost of compounds in inventory will be assessed. Even if compounds are nominally made for a specific project and deposited to a collection, there is the cost of registering the compound, dispensing it, and maintaining the collection. Usually, a company asks its medicinal chemists to make additional compound for inventory when they make the compound for a specific project. Since producing extra compound takes more raw materials and more time for synthesis and purification, there is an inventory cost even for compounds made for a particular project. There is also an opportunity cost associated when a compound is depleted. Unfortunately, complex compounds in inventory that were synthesized for specific projects are seldom resynthesized, since the cost is very large in comparison to the chance that the compound will prove useful.

Replacement cost is one way to estimate the value of inventory. For commercial compounds or chemical library compounds, replacement costs are meaningful. How should the costs of compounds made for a specific project and deposited in inventory to be allocated? A typical cost [cost estimates are based on information provided internally by Glaxo Wellcome chemists in 1993] to make a compound for a specific project is $6000 (costs are highly dependent on current technology[10] and will change over time). The labor to make a typical compound is about $1200, where two-thirds can be attributed to the primary project, one-sixth to current high-throughput screening, HTS (it is common for a large company to have an ongoing general screening effort), and one-sixth to the compound that resides in inventory. The $4500 per compound overhead of running a chemistry department and the $300 per compound materials cost is allocated in the same way. This allocation of cost is summarized in Table 1. We ask our chemists to make 15 mg of compound. A 1 mg amount goes to the primary screen, 3−4 mg goes to current HTS, and 10 mg goes to inventory so that the materials cost is allocated roughly proportionally to where the compound goes. The total inventory cost of the 10 mg is $1150. If 1 mg is needed from inventory for a new screen or a new cycle of HTS, then the compound cost for the sample is about $115 (1150/10).

A typical drug company will have 50−100K or more compounds in its collection. These compounds come from various sources with different costs; estimated costs are given in Table 2. The inventory of a large company will have a value of 50−140 million dollars, or more. An important goal is to maximize the return on this investment. Drug companies typically will screen 10 thousand molecules to find a novel lead molecule with an activity of 1 μM. Synthetic optimization begins with this lead molecule. So the compound cost of getting an initial lead is $600 000−$1 000 000. Only about one-third of the initial leads progress to development. Therefore, three or so initial leads are optimized by synthesizing and testing 300−2000 analogues to progress a compound into development. The target potency of a development compound is typically 1 nM. If it costs $6000 to make an analogue, then the chemical cost of getting a developmental compound is about $4 500 000.

Table 2. Breakdown of Inventory Value, Where M Denotes Millions of Dollars[a]

compound	cost	100K	50K
directed	$115	138.0 M	51.7 M
commercial	$0.02−4	1.2 M	1.6 M
university	$4−6	0.8 M	
total		140.0 M	53.3 M

[a] The cost is for a typical compound or for a collection of 100K or 50K.

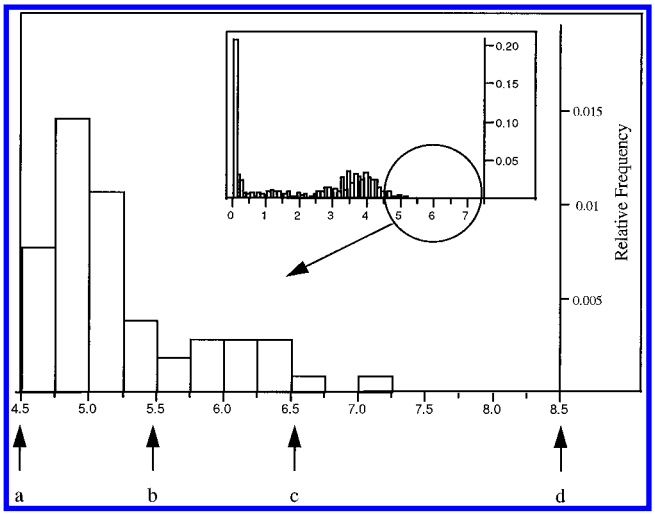

Figure 1. Histogram of the largest 51 observations (lower) comprising the distribution tail which can be seen in the histogram of all potencies (upper right). The letters denote points in the tail of the distribution: (a) where values to the right can be used in statistical estimation, (b) where the percent of compounds to the right is small enough that secondary screening capacity will not be exceeded, (c) where synthetic modification can begin (and HTS cease), and (d) where the project goal is achieved.

EXTREME VALUE ANALYSIS OF SCREENING DATA

The IC_{50} [dose level that results in 50% inhibition] values of 937 compounds [internal Glaxo Wellcome data resulting from screening a large chemical library] were taken for illustrative purposes. The common logarithm (absolute value) of these values is taken to produce the potencies studied. Histograms of the potency distribution appear in Figure 1. The smaller histogram shows the distribution of all potencies, and the larger histogram shows the distribution of the largest potencies. The height of each bar represents the relative frequency of potencies falling in the interval under the bar, where the relative frequency is simply the

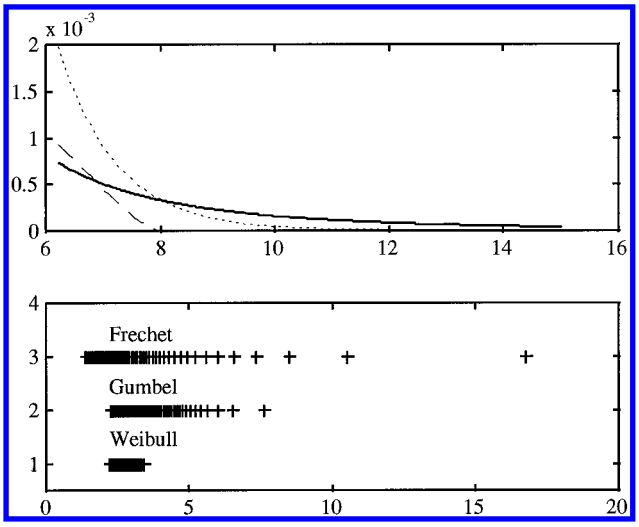

Figure 2. Weibull (dashed), Gumbel (dotted), and Frechet (solid) distribution tails (top). Largest 100 of 1000 predicted observations (bottom). The distributions have been scaled to have the same mean and standard deviations. The Weibull tail has been shifted to the right for comparison of spacings.

number of potencies observed in the interval divided by the total 937. The behavior in the tail of the distribution is of primary concern for predicting large potencies, and, from the histogram, a decision was made to fit the largest 51 observations that appear to make up the right tail of the distribution (see the circled region of Figure 1). There are several points of interest. Various points on the distribution, such as 5.5, 6.5, and 8.5, give guidance for the screening cut-off point, the synthetic modification starting point, and the project goal achievement point. Also, if a mathematical curve can be fit to the right hand tail of the distribution, then the largest values expected in a new larger sample can be predicted. Using extreme value theory, the tail of the distribution can be fit by one of three model distributions. The model then provides relative frequencies of observing very large potencies. Once a distribution is fit to the tail, it is possible to compute the expected value of the largest observations in a future sample.

There are four potency levels to be kept in mind. First, the final target potency of the project, which is typically set around 1 nM or around 8.5 log units in Figure 1. Next, a potency within striking distance for medicinal chemists is set at 2−3 log units below the target potency. This is approximately 6.5, as indicated in Figure 1. Screeners set a somewhat lower potency such as 5.5 for retesting and characterization. All compounds are initially screened at this level, and only for those compounds more potent than this level is a complete dose response curve determined and an IC_{50} estimated. Screeners set their level low enough to provide medicinal chemists with a choice of compounds for synthetic modification and optimization but high enough so that secondary screens are not overwhelmed. Finally, for those compounds with IC_{50}s above the level 4.5 in Figure 1, statisticians can fit a potency distribution and estimate the potencies expected in larger samples.

Fitting the Extreme Value Distribution. One of the three extreme value distributions can be fit to the tail of the empirical potency distribution. The mathematical curves that are fit to the tail of the empirical distribution are completely determined by the relative frequency predictions that they make. Statistically, observations of compound potencies are considered observed values of some random variable X. The proportion (or relative frequency predictions) of potencies smaller than a specified value x, shall be denoted by $\Pr\{X < x\}$ (Distribution Function) and the proportion of potencies larger than x is then $1 - \Pr\{X < x\}$. The parent distribution of compound potency is usually unknown; however, the distribution of the largest observations can still be modeled effectively. Surprisingly, regardless of the parent distribution of the data, for large sample sizes the distribution of the extreme observations usually has a distribution function with one of the follow forms:

type 1 (or Gumbel) distribution:

$$\Pr\{X \le x\} = \exp\left\{-\exp\left(\frac{-(x-\xi)}{\theta}\right)\right\}$$

type 2 (or Frechet) distribution:

$$\Pr\{X \le x\} = \begin{cases} 0, & \text{if } x < \xi \\ \exp\left\{-\left(\frac{(x-\xi)}{\theta}\right)^{-k}\right\}, & \text{if } x \ge \xi \end{cases}$$

type 3 (or Weibull) distribution:

$$\Pr\{X \le x\} = \begin{cases} \exp\left\{-\left(\frac{(\xi-x)}{\theta}\right)^{-k}\right\}, & \text{if } x \le \xi \\ 1, & \text{if } x > \xi \end{cases}$$

This is a key finding from extreme value theory.[11] Hence, these can be viewed as modeling distributions that can be fit to the potency data. The parameters, ξ, θ, and k are chosen to fit the appropriate model distribution to the empirical distribution. Methods for fitting these distributions will be discussed in the next section. First, some guidance is needed in choosing which of the three modeling distributions is appropriate.

A plot of the derivative of the distribution functions with respect to x, which describe the concentration of observations per unit, appear in Figure 2. The derivative of the distribution function (distribution density) is the relative frequency of compounds per unit potency or the likelihood of seeing compounds with a particular potency level. Only the tails of the densities are plotted for comparison. The slower the distribution tail goes to zero (the slower the chances of observing potencies at a level goes to zero as the level increases), the heavier the tail is said to be. The distributions can be ordered by the heaviness of their tails. The Frechet distribution has the heaviest tail and the Weibull distribution the lightest with the Gumbel distribution in between. The heavier the tail, the larger the probability of seeing relatively high extreme values. The spacing of observed potency values provides information on the weight of the tail. It is worth mentioning that the Weibull distribution has a finite upper bound of ξ. There is zero chance of seeing observations larger than ξ when the distribution is Weibull. Since potencies such as IC_{50}s can get arbitrarily close to zero with molecular modification (decreases in orders of magnitude can be observed), the distribution of the PIC_{50}s or IC_{50}s on the common logarithm scale are not likely to follow this distribution.

Distribution quantiles are obtained by finding q_p such that $\Pr\{X \le q_p\} = p$. The Gumbel distribution function,

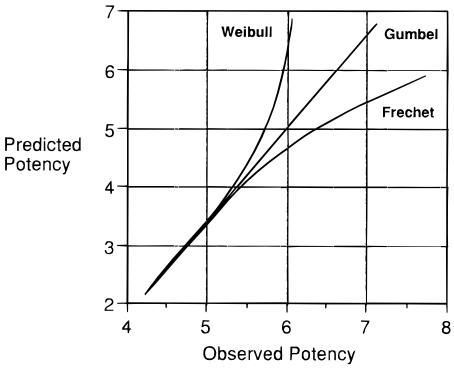

Figure 3. Gumbel probability plot with three distribution types.

evaluated at $x = q_p$, is set equal to p and the equation solved for q_p to produce

$$q_p = -\theta \log(-\log(p)) + \xi \qquad (1)$$

This is the formula for computing the Gumbel quantiles. The quantiles given by (1) predict where the ordered data will fall on average. Hence, the future, ordered observations from a Gumbel distribution can be roughly predicted. The sample quantile, q_{p_i}, where $p_i = (i - {}^1/_2)/n$, $i = 1, 2, ..., n$, is approximately the expected value of the ith largest observation. The quantity p, is the cumulative observed frequency. The parameters θ and ξ are unknown but the fact that the quantiles are a linear function of $-\log(-\log(p))$ means that they can be estimated and used to assess whether or not a set of extremes have a distribution of the Gumbel form. Plotting $-\log(-\log((i - {}^1/_2)/n))$, $i = 1, 2, ..., n$ (the plotting position) against the ordered data (order statistics) produces a linear relationship if the data conform to a Gumbel distribution. The expected spacings of the largest observations can be seen in Figure 2. The Weibull quantiles are closer together than the Gumbel quantiles. As a result, if Gumbel plotting positions are plotted against Weibull quantiles, then they tend to curve upward. In contrast, Frechet quantiles are further apart than Gumbel quantiles. Therefore, if Gumbel plotting positions are plotted against them, then they tend to curve downward. This is illustrated in Figure 3. Formal statistical tests exist for determining the limiting distribution,[12] but they will not be discussed here. Software is available for finding the domain of attraction and for producing the quantile plots. Castillo et al. have produced Macintosh software called Extremes[13] that is quite easy to use. Reiss[14] markets PC software that is powerful, but it is designed for expert statisticians. There is a large amount of literature on fitting the tail of a distribution.[15−17] Using these methods, one can infer which of the three types of tails the distribution has. Since there is good reason to rule out the Weibull distribution, some of these more advanced approaches might produce undesirable results. For example, assay variability or censoring of the largest potencies can lead to the Weibull type tail being inferred. A small simulation study was conducted using draws from a large pool of extreme potencies, and the Weibull distribution was inferred about 20% of the time, where the potencies actually appeared to follow a Frechet or Gumbel distribution.

A Gumbel probability plot, which is restricted to the largest 51 observations, is shown in Figure 4. The points approximately follow a linear relationship that is indicated by the solid line in the figure. The location parameter ξ and

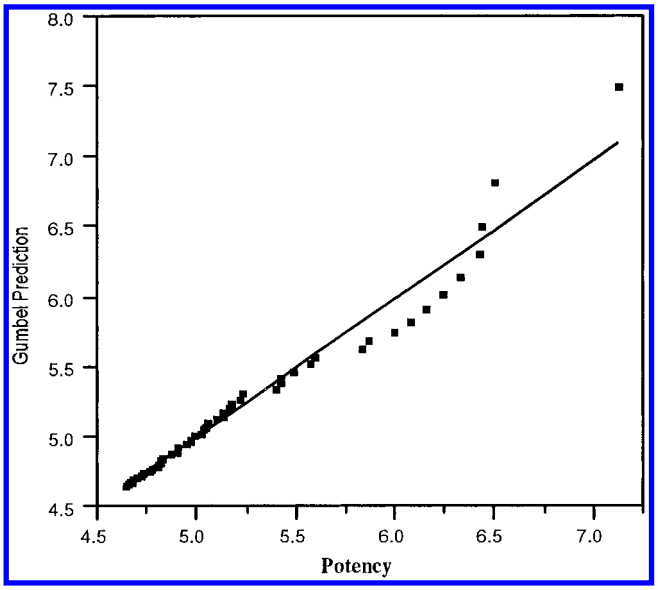

Figure 4. Gumbel plot of largest 51 observations.

the scale parameter θ are chosen to best fit the data using Extremes.[13] The statistical theory and methods for fitting the distributions are quite complex,[15,17−19] but the Extremes software makes solving extreme value problems accessible for those without a strong statistics background. In Extremes, Gumbel plots are done using probability paper and model fitting can be accomplished by one of several methods.

Predicting Highest Potencies in Larger Screening Sets. Methods for estimating the parameters in Extremes include least squares, method of moments, and maximum likelihood.[20] For this analysis, maximum likelihood estimates were chosen. It is of great interest to predict the top k potencies for a new larger screening set (of size n). These will be denoted by $X_{(1)} \geq X_{(2)} \geq ... \geq X_{(k)}$. The strategy is to use a relatively substantial sample to estimate the most potent compounds in a very large collection; for example, screening 10 000 randomly selected compounds to predict the most potent in a 200 000 compound corporate collection. Extrapolation is accomplished by substituting the parameter estimates into the expression below.

$$E[X_{(r)}] = \xi + \theta\left\{\frac{n!}{(r-1)!(n-r)!}\sum_{j=0}^{r-1}(-1)^j\binom{r-1}{j} \times \left[\frac{\gamma + \log(n+1-r+j)}{(n+1-r+j)}\right]\right\} \qquad (2)$$

This is the expected (or average) value of the rth largest observation assuming a Gumbel distribution. Parameter estimates will be indicated by the caret overmark and the rth largest observation prediction will be denoted by $\hat{X}_{(r)}$. Denoting the term in curly braces by c_r makes the right hand side of (2) equal $\xi + \theta c_r$. It is true that

$$\mathrm{Std}[\hat{X}_{(r)}] = [\hat{\sigma}_{\hat{\xi}}{}^2 + c_r{}^2\hat{\sigma}_{\hat{\theta}}{}^2 + 2c_r\hat{\sigma}_{\hat{\xi}\hat{\theta}}]^{1/2} \qquad (3)$$

where Std denotes the standard error of the prediction, $\hat{\sigma}_{\hat{\xi}}{}^2$, denotes the variability of the parameter estimate ξ, $\hat{\sigma}_{\hat{\theta}}{}^2$ denotes the variability of the parameter estimate θ, $\hat{\sigma}_{\hat{\xi}\hat{\theta}}$ denotes the covariance of the two parameters, and $\gamma = 0.577\,22$ denotes Euler's constant. These are all quantities produced by Extremes when using maximum likelihood

Table 3. Predictions of the Largest Five Potencies Expected for Several Screening Set Sizes, n[a]

n	$X_{(5)}$	$X_{(4)}$	$X_{(3)}$	$X_{(2)}$	$X_{(1)}$
937	6.16	6.31	6.51	6.82	7.44
(std err)	(0.76)	(0.72)	(0.74)	(0.79)	(0.87)
(obsd)	(6.33)	(6.43)	(6.44)	(6.50)	(7.13)
2000	6.62	6.78	6.98	7.29	7.90
(std err)	(0.76)	(0.78)	(0.81)	(0.85)	(0.93)
5000	7.19	7.34	7.54	7.85	8.47
(std err)	(0.84)	(0.86)	(0.88)	(0.93)	(1.01)
10K	7.61	7.77	7.97	8.28	8.89
(std err)	(0.89)	(0.91)	(0.94)	(0.98)	(1.07)
20K	8.04	8.19	8.40	8.70	9.32
(std err)	(0.95)	(0.97)	(1.00)	(1.04)	(1.12)
50K	8.60	8.75	8.96	9.27	9.88
(std err)	(1.03)	(1.05)	(1.08)	(1.12)	(1.20)
100K	9.03	9.18	9.38	9.69	10.31
(std err)	(1.09)	(1.10)	(1.13)	(1.18)	(1.26)

[a] The potency measurements are on the log scale.

Table 4. Cost in Millions of Dollars of Screening 2000 Compounds as a Function of the Number of Screens and the Amount of Compounds Used per Screen[a]

	5 screens	10 screens	20 screens	40 screens
1.0 mg	1.2	2.3	4.6	9.2
0.6 mg	0.7	1.4	2.8	5.5
0.2 mg	0.2	0.5	0.9	1.8

[a] Costs are based on the figure of $115/mg.

estimation. SAS code is available from the authors for computing (2) and (3).

The following maximum likelihood estimates where found using the 51 largest potencies:

$$\hat{\xi} = 2.88 \qquad \hat{\theta} = 0.614 \qquad \hat{\sigma}_{\xi}^{2} = 0.0664$$

$$\hat{\sigma}_{\theta}^{2} = 0.007\,02 \qquad \hat{\sigma}_{\xi\theta}^{2} = 0.0203$$

The fitted probability model can now be used for predicting the most potent compounds in a screening sample of any given size. The five largest potencies that can be expected for various screening set sizes are predicted by using the parameter estimates along with relations 2 and 3. The potency predictions for the screening data appear in Table 3. Potencies are on the log scale, and it can be seen that a 5-fold increase in the number of compounds screened is required to obtain a 10-fold increase (1 log unit) in potency.

DETERMINING OPTIMAL SCREENING STRATEGIES

The cost of screening is determined by how many compounds are tested, the number of screens, and the amount of compound tested in a screen. If a company screens 2000 compounds in an assay, then Table 4 contains the cost in millions of dollars as a function of the amount of compound per screen and the number of screens. A 5-fold increase in the number of compounds (10 000) screened will result in five times the costs.

At some point synthetic modification is more cost effective than screening. It requires a large multiplicative increase in the number of random compounds screened in order to get a log unit increase in potency. We assumed that synthetic modification will yield a log unit increase in potency, for a fixed number of analogues in a linear fashion; our experience is that 1 log increase in potency takes about 100 analogues.

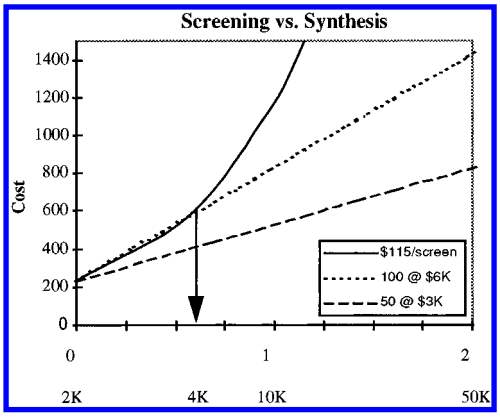

Figure 5. Cost of improvement in thousands of dollars plotted against both the increase in potency on the log scale and the number of compounds required to get the increase in potency by screening. The solid line is the cost based on screening, and the other curves are based on a fixed number of analogues per log unit potency increase. The screening cost is based on $115 per compound.

The number of analogues required to reach potencies at the development level will vary from project to project. The cost of creating analogues is much higher than the cost of compounds already available from inventory, but the increase in potency per compound that can be expected is better. Figure 5 shows the break-even point between screening and synthetic modification. The y-axis gives the cost of improvement. The x-axis is indexed with two scales. First, potency is given in log units, 0, 1, 2; 1×, 10×, and 100×. The second scale is the number of compounds necessary to get that increase in potency by screening. The curved line relates the cost of screening to the expected increase in potency. The two lines give the expected increase in potency when synthetic analogues are made using two different cost per number-of-analogues assumptions. The dotted line is based on the assumption that it takes 100 synthetic analogues to get a log unit increase in potency with a cost of $6000 per compound. The dashed line is based on the assumption that it takes 50 synthetic analogues at $3000 a piece to get a log unit increase in potency.

If making analogues is inexpensive and medicinal chemists expect to get a log increase in potency with 50 compounds, then they should immediately switch to synthesis. However, if synthesis is expensive and it is expected to take 100 analogues to get a 1 log increase in potency, then the cost break-even point is about 4000 compounds screened.

The amount of compound required for performing a screen has a large impact on the cost of screening. Suppose the cost per screen is cut roughly in half. The break-even points under the assumption of a cost of $60 per screen are shown in Figure 6. When the cost of synthetic modification is low (50 compounds per log unit increase at $3000 per compound), the screening cost and synthetic modification cost are nearly identical until about 3000 compounds have been screened. At this point synthetic modification is much cheaper.

However, if the cost of synthetic modification is high (100 compounds per log unit increase at $6000 per compound), then it is more cost effective to screen as many as 16 000 compounds before switching to synthetic modification. A similar graphical analysis shows that if the cost of screening is as low as $11.5 per screen, then it is economically feasible to screen over 1.4 million compounds before beginning synthetic modification, when the cost of synthetic modifica-

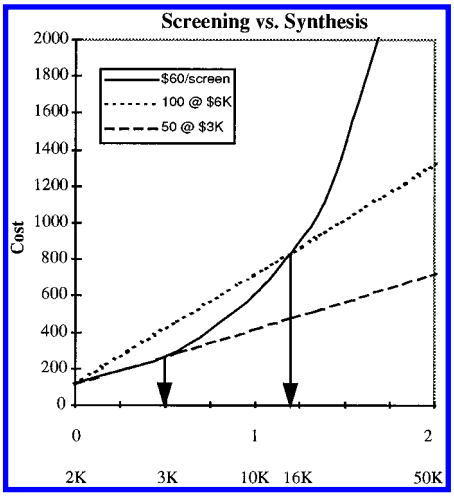

Figure 6. Cost of improvement in thousands of dollars plotted against both the increase in potency on the log scale and the number of compounds required to get the increase in potency by screening. The solid line is the cost based on screening, and the other curves are based on a fixed number of analogues per log unit potency increase.

tion is high. This would quite likely require using all the compounds available in inventory. The break-even point between screening and synthesis is sensitive to compound cost. If combinatorial synthesis can bring the cost down, then it will make economic sense to screen more compounds.

PREDICTING THE FREQUENCY OF POTENCIES EXCEEDING A LARGE VALUE

Given a potency level of interest, such as one of the labeled points on Figure 1, the relative frequency at which the point will be exceeded is predicted by $1 - \Pr\{X < x\}$, where x is the point of interest and $\Pr\{X < x\}$ is given by one of the three extreme value distributions. The relative frequency of an initial sample can be used to predict the number of compounds, k, which will have potencies exceeding x. The predicted number of compounds is given by $k = N(1 - \Pr\{X < x\})$, where N is the total number of compounds screened.

The tail observations have been used to fit the extreme value distribution so far. However, another option is to arbitrarily group the compounds into sets of size q, where q could be 100 or 250, and find the largest potency in each group. One of the extreme value distributions can then be fit to these group maximums with Extremes. This is commonly done in the analysis of maximum flood levels[21] or maximum temperatures.[22] The yearly maximums are used. The extreme value model can be used to predict the largest potencies of compounds that will be observed in a sample of size $n = mq$. Formula 2 must be applied with m in place of n. In cases where the Gumbel plot of the tail observations does not follow one of the three forms shown in Figure 3, grouping is an alternative course of action. Grouping can be done in such a way to put very similar compounds into the same group in order to cut down on serial correlations that might make the extreme value analysis less reliable. This could be done by ordering the compounds by a similarity index[23] before grouping them consecutively. Statistical theory on the validity of extreme value analysis in this context exists.[11]

If high potencies of several very similar compounds are a part of the data used to fit the extreme value distribution, then the predictions will be inaccurate. However, grouping the compounds arbitrarily and taking extremes makes poorer use of the data than fitting the tail observations. Paying careful attention to the similarity of the structures of compounds with high potencies might be a better alternative than just arbitrarily grouping the data. The potencies of compounds with similar structures could be replaced by the potency of the best molecule of the similarity set.

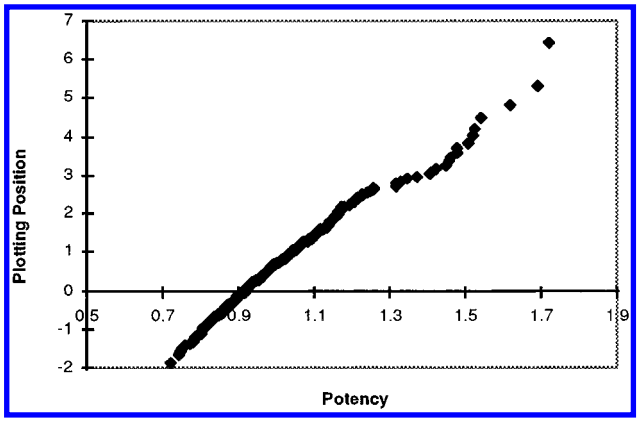

Figure 7. Gumbel plot for the single point assay values.

Validation of the Extreme Value Theory Methods. The typical way to estimate the relative frequency of compounds exceeding a threshold in a large sample is to count the number of observations that exceed the threshold in a small sample. The fraction of the total number of compounds with potencies exceeding the threshold is then computed. The use of extreme value theory is typically better than this simple count. It is clear that in a small sample there might be no observations that exceed a threshold, whereas the formulas can be used to estimate the fraction expected to exceed the threshold ($1 - \Pr\{X < x\}$), and this estimate might be more reliable.

To evaluate the fit of an extreme value distribution to an observed distribution, a large data set is required. Large data sets of single point assay results are more available, but more variable than PIC_{50} values. (Typically, single point assay results are expanded for the most potent compounds so that PIC_{50}s are available for the extreme values. Where available, PIC_{50}s should be used.) We used single point assay results for 76 500 compounds [internal Glaxo Wellcome data resulting from screening a large number of available compounds in a cell based assay] in a simulation to demonstrate the greater stability of theoretical estimates over the simple count of observations over a threshold. To decrease the effect of serial correlation with the compounds, the 76 500 compounds were grouped consecutively into 306 groups of size 250 and the largest potency value of each group was recorded. A Gumbel plot of the group extremes is given in Figure 7. The distribution of these extremes looks to be either Frechet or Gumbel.

Next, random samples of size 30 were taken from the 306 group extreme values. Approximately 10% of the total was chosen since it supplied an ample amount of data for estimation, yet the percentage of the total is small enough that the benefit of the extreme value distributions can be clearly seen. The count method will do better in comparison as the proportion of the sample used for prediction increases.

For the observed distribution of the 30 potencies, a threshold of 1.5 was selected. This value was considered

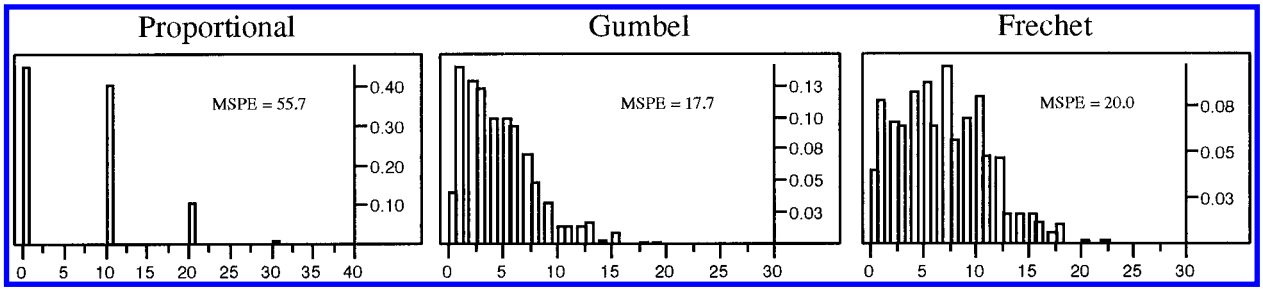

Figure 8. Simulation results. Mean squared prediction error (MSPE) with relative frequency histograms (relative frequency on the vertical axes) of the predicted number of compound potencies exceeding the threshold based on 500 random samples of 30 from the 306 group maximums. The true number of potencies exceeding the threshold is 7.

of practical importance. There are seven extreme values that exceed 1.5 among the 306 values. For each set of 30 randomly selected extreme values, the proportion that exceeds 1.5 was estimated by a simple count and by fitting the Frechet and Gumbel distributions. For instance, if 2 out of the 30 randomly selected extremes exceeded 1.5, then the count estimate of the proportion would be 0.067 (2/30). The extreme value distributions were fit to the distribution of the 30 extremes by standard methods discussed in the references.[19]

A summary of the results is given in Figure 8. Histograms are used to show how the predictions using each of the three methods are distributed. We know that in the full sample of 306 extreme values that seven exceed 1.5. Using the simple count, in 45.8% of the samples of 30, the estimate of the number of compounds with potencies exceeding 1.5 was zero, whereas the theory formulas gave estimates of zero only about 4% of the time. About 65% (62% Gumbel and 66.4% Frechet) of the time the formulas gave estimates of the number of compounds with potencies exceeding 1.5 to within ± 4 of the correct answer of 7, whereas this occurred only 38.6% of the time with the count. The mean square prediction error, MSPE, is the square of the difference between the prediction based on a sample and the known value of 7 averaged over all of the random samples of 30. We want this prediction error to be small, and it is much smaller for the formulas than for the simple counts. So the formula predictions are much less likely to produce a gross underestimate of the true value and also more likely to produce an estimate close to the true value. Both the Frechet and Gumbel predictions perform about the same for this data set. Since the Gumbel distribution is easier to fit, it might be preferred. To estimate the parameters θ and σ for the Gumbel distribution, some software package such as Extremes should be used. The Frechet distribution is harder to fit. Extremes requires θ to be specified for this distribution. The smallest potency can be used, but this value can be quite far from that selected by a statistically sound method such as the maximum likelihood. In order to fit the parameter θ statistically, the PC software Xtremes[14] must be used or some other software package. Once the parameters are available, they can be substituted into the appropriate extreme value distribution formula to obtain the predicted proportion, $1 - \Pr\{X < 1.5\}$.

DISCUSSION

Inventory is valuable. We show that linear improvements from screening require geometrically more compounds. Optimal screening set size is determined by inventory costs and the cost of gain from synthetic modification. The optimal screening set size is very sensitive to the cost per screen and the cost of expected gain from synthetic analogues.

In the past, it is likely that the number of compounds screened in a project has been determined rather arbitrarily by how many compounds were available and the amount of time that could be spent on screening. Certainly, it would be better to use resources optimally. The framework presented here can be used for deciding how many compounds to screen in a program. The answers depend on costs, so it will pay to obtain accurate cost data. The potencies of the most potent compounds in a collection can be estimated if good sample data are available. The accuracy of the inference in the case of a one shot assay was assessed and shown to be more informative than just looking at whether there were any compounds that exceeded a threshold potency value of interest in an exploratory screen. The extrapolations have a high degree of variability, but they can be used as rough guides for future screening. For example, if random samples from two compound collections are made and if we know the size of the two collections, then these methods can be used to suggest which collection is more likely to have the better compounds.

This analysis was performed assuming that one random sampling of the collection would be done (with a possible second random sample) followed by a switch to analogue synthesis. In practice, some sequential sampling is usually more cost effective. For example, after the initial sampling, the most potent compounds can be used to drive a similarity search to choose a new screening set with an expected higher concentration of potent molecules. This could reduce the cost of screening in the second stage.

CODE AVAILABILITY

SAS code is available from M.F. for extreme value calculations, formulas 2 and 3 and for related equations for Frechet based predictions.

ACKNOWLEDGMENT

Mike Hayes and Steve Blanchard provided much of the cost information. E. Castillo made the Extremes package available and answered many questions about its use. Richard Smith answered some tough extreme value theory questions. We are also grateful to Glaxo Wellcome for supporting this work.

REFERENCES AND NOTES

(1) McFarland, J. W. *Chronicles of Drug Discovery*; Bindra, J. S., Lednicer, D., Eds.; John Wiley and Sons: New York, 1983; Vol. 2, pp 87−108.

(2) Geysen, H. M.; Meloen, R. H.; Barteling, S. J. Use of Peptide Synthesis to Probe Viral Antigens for Epitopes to a Resolution of a Single Amino Acid. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 3998−4002.
(3) Bunin, B. A.; Ellman, J. A. A General and Expedient Method for the Solid-Phase Synthesis of 1,4-Benzodiazepine Derivatives. *J. Am. Chem. Soc.* **1992**, *114*, 10997−10998.
(4) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Foder, P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1400.
(5) Chaiken, I. M., Kim, D. J., Eds. *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; American Chemical Society: Washington, DC, 1996.
(6) Broach, J. R.; Thorner, J. High-throughput Screening for Drug Discovery. *Nature* **1996**, *364* (Suppl), 14−16.
(7) Willett, P. *Similarity and Clustering in Chemical Information Systems*; John Wiley and Sons: New York, 1987.
(8) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(9) Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Method on an Initial Sample. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 66−71.
(10) *Pharmaceutical R&D: Costs, Risks and Rewards*; Office of Technology Assessment, U.S. Congress: Washington, DC, 1993; pp 105−134.
(11) Leadbetter, R. L.; Lindgren, G.; Rootzén, H. *Extremes and Related Properties of Random Sequences and Processes*; Springer-Verlag: Berlin, 1982; Chapter 1, p 4.
(12) Hill, B. M. A Simple Approach to Inference about the Tail of a Distribution. *Ann. Stat.* **1975**, *3*, 1163−1174.
(13) Castillo, E.; Alvarez, A.; Cobo, A.; Herrero, M. T. *An Expert System for the Analysis of Extreme Value Problems*; Department of Applied Mathematics, University of Cantabria: Cantabria, Spain, 1993.
(14) Reiss, R. D.; Hassmann; Thomas, M. *Xtremes: Extreme value analysis and robustness*; Xtremes Group Seigen, University of Seigen: Seigen, Germany, 1996.
(15) Davison, A. C.; Smith, R. L. Models for Exceedances over High Thresholds. *J. R. Stat. Soc. B* **1990**, *52*, 393−442.
(16) Hosking, J. R.; Wallis, J. R. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics* **1987**, *29*, 339−349.
(17) Weissman, I. Estimation of Parameters and Large Quantiles Based on the k Largest Observations. *J. Am. Stat. Assoc.* **1978**, *73*, 812−815.
(18) Castillo, E. *Extremes in Engineering Applications*. Conference on Extreme Value Theory and Its Applications, National Institute of Standards and Technology. *J. Res. Natl. Inst. Stand. Technol.* **1994** (Spring).
(19) Hosking, J. R.; Wallis, J. R.; Wood, E. F. Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics* **1985**, *27*, 251−261.
(20) Castillo, E. *Extreme Value Theory in Engineering*; Academic Press: New York, 1988; Chapters 4 and 5.
(21) *Flood Studies Report*; Natural Environment Research Council: London, 1975; Vol. 1.
(22) Jenkinson, A. F. The Frequency Distribution of the Annual Maximum (or Minimum) of Meteorological Elements. *Q. J. R. Meteorol. Soc.* **1955**, *81*, 158−171.
(23) Burden, F. R. Molecular Identification Numbers for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255−257.

CI970224+