JOURNAL OF combinatorial CHEMISTRY

Subscriber access provided by American Chemical Society

Article

Confirmation of Primary Active Substances from High Throughput Screening of Chemical and Biological Populations: A Statistical Approach and Practical Considerations

Ji-Hu Zhang, Thomas D.Y. Chung, and Kevin R. Oldenburg

J. Comb. Chem., **2000**, 2 (3), 258-265• DOI: 10.1021/cc9900706 • Publication Date (Web): 09 April 2000

Downloaded from http://pubs.acs.org on March 20, 2009

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

ACS Publications
High quality. High impact.

Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Confirmation of Primary Active Substances from High Throughput Screening of Chemical and Biological Populations: A Statistical Approach and Practical Considerations

Ji-Hu Zhang,* Thomas D.Y. Chung, and Kevin R. Oldenburg

Leads Discovery, Experimental Station, DuPont Pharmaceuticals Company, P.O. Box 80400, Wilmington, Delaware 19880-0400

Received November 3, 1999

Many biologically important substances are discovered through screening of relevant chemical or biological libraries. The ability to find the active substances ("hits") from any random collection is largely determined by the quality of the assay and screening conditions. When a large population is screened for a specific characteristic, each member of that population is usually tested only once. Errors in the measurements require additional follow-up tests to confirm that each hit from the primary screen is truly *active*. In this report, we present a statistical model system that predicts the reliability of hits from a primary test as affected by the error in the assay and the choice of the hit threshold (*hit limit*). The hit confirmation rate, as well as false positive (representing substances that initially fall above the hit limit but whose true activity are below the hit limit) and false negative (representing substances that initially fall below the hit limit but whose true activity are in fact greater than the hit limit) rates have been analyzed with this model by computational simulation. This model can also be used in screen validation and post-screening data analysis. The statistical analysis presented here has broad implications and is applicable to screening of any large population for any specific characteristic. Obvious applications include drug discovery, gene chip analysis, population biology, directed molecular evolution, biological panning, and combinatorial material sciences.

Introduction

Screening of random compound collections for identification of leads for drug development has been largely a practical endeavor in the pharmaceutical industry for many years.[1,2] Recent advances in drug target identification,[3−5] chemical compound library construction,[6,7] development of high throughput assay methodologies,[8] instrumentation, automation, and ADME (adsorption, distribution, metabolism, and excretion) characterization have transformed the area of high throughput drug screening into a rapidly growing, multi-disciplinary scientific and technological field. As the throughput of screening increases and the strategies for construction of chemical libraries is optimized, it is almost certain that the number of active compounds ("*hits*") resulting from primary screening will increase considerably. The efforts to confirm the activity of these increasing numbers of primary hits, either from a potency screen or an ADME-type screen, will be nontrivial.

In most high throughput screening (HTS) programs, each substance is usually tested only in singlet due to reagent, labor, time, and cost considerations. Hits need to be identified in the presence of and despite a certain degree of variability in signal measurement from any assay. A high degree of accuracy and sensitivity in the assay is therefore critical for identifying hits. A *high quality* HTS assay must be able to identify, with a high degree of confidence, those few substances that display (biological) activity significantly different from the rest of the chemical or biological library. One of the fundamental issues that needs to be clarified is the reliability (or confirmation) of screening results and the main factors affecting it. Moreover, since each HTS campaign is a significant expenditure in terms of labor, time, and resources, it is crucial to have, during the assay and screen validation process, a quantitative analysis of the capability of a particular screen to identify the majority of potential hits residing in the chemical or biological library. The confirmation rate of hits from the primary screen is mainly affected by (a) the HTS assay quality, (b) the hit limit selection, and (c) the primary hit profile. Among them, the assay quality plays the primary role in determining the hit confirmation rate.[9] The scope of this paper is to further delineate the effects of these factors on the primary hit confirmation rate as well as on the false positive and false negative rates, from both a statistical and a practical view.

The drug discovery analysis presented here is a specific case from the broad spectrum of possible applications of this statistical model. There are obvious implications in areas such as population biology, directed molecular evolution, biological panning, gene chip analysis, and combinatorial material sciences.

A Statistical Model

Screening a large, random chemical or biological population in order to find the few active substances (hits) involves

* To whom correspondence should be addressed.

10.1021/cc9900706 CCC: $19.00 © 2000 American Chemical Society
Published on Web 04/09/2000

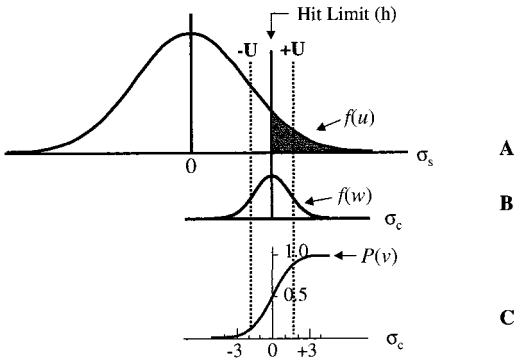

Figure 1. Illustrations of the statistical distribution functions used in the text: A, the population distribution, $f(u)$, in respect to the hit limit (h) and primary hits (shaded area). Function $f(u)$ a priori can be any population distribution. A Gaussian distribution is assumed here for $f(u)$ with a standard deviation, σ_s. B, the measurement variability function, $f(w)$, which is a Gaussian distribution with a standard deviation, σ_c. C, the corresponding cumulative Gaussian distribution of $f(w)$, namely $P(v)$, which represents the confirmation probability of a compound at a specific activity (shown one at $+U$ and one at $-U$) in respect to the hit limit (h).

statistical decisions. These statistical decisions are related to classical hypothesis testing used in quality control evaluation.[10] For drug screening, a substance (compound) scores as "active" when its activity lies *at* or *beyond* the set threshold (i.e., the substance shows an activity different from the mean activity of the population by the threshold value). The threshold (the *hit limit*) for declaring a *hit* (usually an activity outlier) is generally expressed as a certain activity value or a number of standard deviations away from the mean of the library population in a primary screen. To be consistent and avoid any discrepancy in derivation and use of the working model and its conclusions, the term "confirmation" of hits is specified here as the following.

Confirmation. Upon retesting, a *confirmed* hit is defined as having an activity *equal to or greater than* the threshold value (the hit limit) by which the primary hits were identified in the primary screen.

Thus, *the hit confirmation rate* refers to the ratio of the number of confirmed hits to the total number of primary hits included in the confirmation testing from a primary screen. Figure 1 illustrates the population distribution and the hit confirmation probability. Clearly, the primary hits are the fraction of the population to the right of the hit limit, h (shaded area in Figure 1A. See Appendix for an explanation of each symbol or abbreviation used in the text).

At activity U, the confirmation rate is proportional to the probability of a hit being confirmed[9] multiplied by the frequency (or density) of compounds at that activity. Therefore, the overall confirmation rate (CR) for all the primary hits ($u \geq h$) from a population screen is

$$\mathrm{CR} = \frac{\int_h^\infty f(u)\, P(v)\, \mathrm{d}u}{\int_h^\infty f(u)\, \mathrm{d}u} \tag{1}$$

where $P(v)$ is the probability function for hit confirmation (Figure 1C, which is related to the variability of measurements in the screen, Figure 1B) and $f(u)$ is the distribution density function (or distribution of population, Figure 1A). $f(u)$ in eq 1 can be any function that approximates the population distribution. The numerator integral in eq 1 represents the total confirmed hits, and the denominator integral corresponds to the total number of primary hits.

The probability of false positives and false negatives can also be assessed. *False positive hits* refer to those primary hits that are subsequently shown to have an actual activity below the hit limit. The overall *false positive rate* (FPR) is given by

$$\mathrm{FPR} = 1 - \mathrm{CR} \tag{2}$$

False negative hits are those that are missed in a primary screen. The *false negative rate* (FNR) of a screen is defined as the following:

FNR = number of missed hits/(number of missed hits + number of confirmed hits).

The expression for calculating the false negative hit rate is thus given by

$$\mathrm{FNR} = \frac{\int_{-\infty}^h f(u)\, P(v)\, \mathrm{d}u}{\int_{-\infty}^h f(u)\, P(v)\, \mathrm{d}u + \int_h^\infty f(u)\, P(v)\, \mathrm{d}u} \tag{3}$$

where again, $P(v)$ is the probability function for hit confirmation and $f(u)$ is the distribution density function (distribution of population, Figure 1). It should be noted here that the confirmation, false positive and false negative hit rates defined above are conventional in drug screening[1] and may be distinct from those used in statistical hypothesis testing.[10] In the same vein, statistical "outliers" may not be the only hits in conventional drug screening.

A Case Study: A Gaussian Population Approach for Confirmation, False Positive, and False Negative Rates

To deduce a useful statistical model for the hit confirmation rate, it is first assumed that each individual hit, when tested repeatedly, regardless of its activity, exhibits a Gaussian distribution profile around its "true" value of activity. Second it is assumed that the error at any point in the activity dynamic range is kept the same, with a *constant* standard deviation, σ_c (the constant error assumption). The latter assumption is for the simplicity of the model and is obviously not always true since in many cases the errors are "proportional" (vide infra).

Under these assumptions, the probability function for hit confirmation, $P(v)$, in eqs 1 and 3 can be expressed as the cumulative function of a Gaussian distribution function for the variability of measurements, $f(w)$, which is closely related to the error function (Figure 1B). It should be noted that $P(v)$ is the same function used for the probability of a substance being declared a hit (scored "positive") in primary screening.[9,11]

Different chemical and biological libraries will have different population distribution profiles. As a case study, it is further assumed here that the (chemical or biological) population approximates a Gaussian or normal distribution. This is mainly for the theoretical calculations as in many practical cases the population distribution will deviate from

Gaussian. (The distribution of a particular variable of a population may deviate at various extents from an ideal Gaussian distribution in modality, skewness, kurtosis, etc.) However, for the primary screen based on a large, unbiased, random chemical or biological library, the activity profiles from these chemical or biological populations would approximate a Gaussian or normal distribution based on the central limit theorem. This assumption largely holds true for at least the bulk of the population of random collections.[9,11] Under these limitations, the formula for the population distribution, $f(u)$, can be estimated by a Gaussian distribution function, with a standard deviation of σ_s. Therefore, σ_s contains both the variability between different members of the library (i.e., treatment effect or diversity) and the measurement error of the assay.

Under the above assumptions, the hit confirmation rate (CR), the false positive rate (FPR), and the false negative rate (FNR) can be expressed as functions of σ_c (the standard deviation of the Gaussian function for measurement variability), σ_s (the standard deviation of the bulk population distribution, this representing a combination of treatment effects and assay measurement error), and h (the hit limit cutoff value). The detailed expression equations and their deduction are given in the Appendix of this article. From these equations (eqs A4 and A5 in Appendix), the CR and FNR values at the extremes can be deduced. For example, when the assay has no measurement error, i.e., when $\sigma_c \rightarrow 0$, $P(v) \rightarrow 1$ (for activity $U \geq h$) or $P(v) \rightarrow 0$ (for $U < h$). This is equivalent to $P(v)$ being a step function at h. Thus, CR $\rightarrow$ 1 and FNR $\rightarrow$ 0. These results are obvious because when there is no error in measurement, every primary hit should be a true hit and no hit is missed from the screen. At the other extreme, when σ_c is extremely large ($\sigma_c \rightarrow \infty$), then $P(v) \rightarrow 0.5$, and CR $\rightarrow$ 0.5. All the declared hits have a 50−50 chance of being confirmed. CR = 0.5 is therefore the lower limit for the hit confirmation rate. FNR under this extreme condition depends solely on the hit limit setting.

The hit limit setting not only affects the number of declared hits from a given screen, but also affects the number of false-positive and false-negative hits. Basically, the hit limit setting (h) affects the hit confirmation rate and the false negative rate by changing the integration boundaries of eqs 1 and 3. It should be noted that eq 1 defines the *overall* hit confirmation rate ($u \geq h$) of a Gaussian population and can be easily modified for calculating the hit confirmation rate within any given range of activity from a to b ($b > a \geq h$), by adapting these boundaries to the integrals in the equation. Most notably, it will give the confirmation rate of a specific "bin" of primary hits.

Equations 1 and 3 above are derived for a screen where each compound is measured only once. If multiple independent measurements are performed for each data point, then the standard deviation of the mean of n measurements is given by $\sigma_{ave} = \sigma_c/\sqrt{n}$. Use of σ_{ave} in place of σ_c in the equations will yield the confirmation rate for multiple measurements (see Appendix).

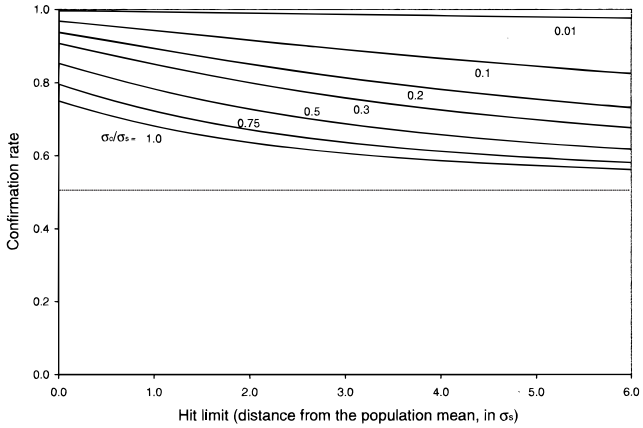

Figure 2. Calculated confirmation rate curves based on eq 1 in the text. The *x*-axis is the distance of the hit limit from the mean activity of the population, in standard deviations (σ_s). σ_s contains both the assay measurement error and the variability between each members of the library (i.e., treatment effect). Each line represents a specific assay variability ratio (σ_c /σ_s) as indicated. The curve beyond $3\sigma_s$ is shown mainly for displaying the trend of the curve for a pure Gaussian distribution, since any hit in this range ($>3\sigma_s$) is likely to belong to a second population distribution. When the hit limit is set infinitely far to the right from the population mean, all the curves will approach CR = 0.5 (shown as the horizontal dotted line). The calculation is performed using the IMSL QDAG univarate quadrature routine (see Appendix).

Effects of Hit Limit Setting and Screening Assay Quality

Figure 2 shows a series of curves that simulate the overall confirmation rate (see eq A4 in Appendix) as a function of the distance that the hit limit is placed from the mean activity of the population (μ_s), in units of standard deviation (σ_s), based on a Gaussian population. Numerical integration was performed using the IMSL QDAG univariate quadrature routine (ref 12, see Experimental Protocol in Appendix). The results from this calculation demonstrate that the overall confirmation rate of the primary hits varies with the placement of the hit limit and the assay quality. Intuitively, if only potent hits were chosen from the primary screen, the likelihood of these hits being confirmed is high. Somewhat surprisingly, the general trend is that the overall confirmation rate decreases as the hit limit (h) moves further away from the mean activity of the Gaussian population (μ_s). The decrease in the confirmation rate is more prominent in the early part of the curve and gradually levels off as the hit limit is further away from the activity mean of the population. For any screen with a finite measurement error (σ_c), the confirmation rate approaches 0.5 when the hit limit is set infinitely far from the mean of the population. Consequently, setting the hit limit further away from the activity mean of the Gaussian population will yield hits with higher activity (at higher confidence level for being "outliers") but will decrease the hit confirmation rate. The reasons for this seemingly counter-intuitive result are severalfold. First, as mentioned earlier, the definition of a "confirmed hit" is to compare the hit's activity from confirmation testing to the pre-set "hit limit" value in the primary screen rather than to the population activity mean. Second, this result is based on a pure Gaussian population model. When the normality of a real population deviates from pure Gaussian, which is usually

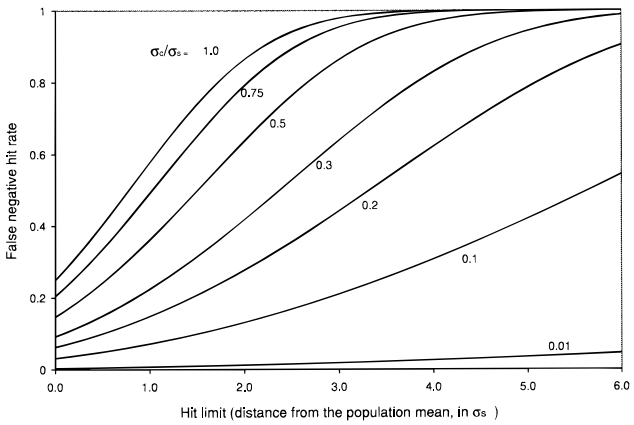

Figure 3. Calculated false negative rate curves based on eq 4 in the text. The *x*-axis is the distance of the hit limit from the mean activity of the population, in standard deviations (σ_s). σ_s contains both the assay measurement error and the variability between each members of the library (i.e., treatment effect). Each line represents a specific assay variability ratio (σ_c/σ_s) as indicated. The curve beyond $3\sigma_s$ is shown mainly for displaying the trend of the curve for a pure Gaussian distribution, since any hit in this range ($>3\sigma_s$) is likely to belong to a second population distribution.

the case, the confirmation rate prediction in these cases would deviate from the model case, particularly when the hit limit is set beyond $3\sigma_s$ from the population mean, as will be discussed further in the next sections.

It is important to note that the confirmation rate depends heavily on the assay variability, σ_c. The confirmation rate improves significantly as σ_c decreases. For example, at a hit limit set at $3\sigma_s$, an assay with variability (σ_c) values of $1.0\sigma_s$, $0.5\sigma_s$, and $0.2\sigma_s$ will yield overall hit confirmation rates of 0.60, 0.69, and 0.81, respectively (Figure 2). The assay variability also determines the lower hit limit setting. The lower hit limit is the lower limit of detection (LLD) of the assay, i.e., at $h \geq 3\sigma_c$ above the mean activity (which is equal or close to zero). These results clearly indicate that the assay data variability plays an important part in determining the confidence level of the primary hits generated from the screen. Recently it has been suggested that the assay quality can be expressed by the Z'-factor[9] ($Z' = 1 - (3\sigma_{c+} + 3\sigma_{c-})/|\mu_{c+} - \mu_{c-}|$, see Appendix for Z'-factor and Z-factor definitions.). It can be seen that, under the above constant error assumption (such that σ_{c+} and σ_{c-} are equal or comparable to σ_c), the Z'-factor is inversely proportional to σ_c, assuming the dynamic range of the assay is kept constant. Therefore, the assay quality (the Z'-factor) has a direct effect on the confirmation rate when the assay is used for screening. It is interesting to note that the confirmation rate has a direct relationship with the Z'-factor of the assay, not necessarily with the Z-factor of the screen. The reason for this is that *the Z-factor for a screen is more indicative of the screen "window" for identifying hits and is test library sensitive while the Z′-factor is indicative of the measurement variability of the assay and is test library independent*.[9]

On the basis of eq 2, the curves that describe the false positive hit rate (FPR) are therefore the mirror images (about the line CR = 0.5 in Figure 2) of those for the confirmation rate. Figure 3 shows the calculated false negative hit rate as a function of the hit limit and assay quality (see Experimental Protocol in Appendix for calculation details). It can be seen that, as the assay quality deteriorates (σ_c becomes larger), the false negative rate increases sharply then levels off as it approaches a maximum value of 1. For example, when $\sigma_c = 0.3\sigma_s$ (the assay variation (σ_c) is about 30% of the standard deviation of the Gaussian population (σ_s)) and the hit limit is set at $2.5\sigma_s$ from the mean, the assay will miss approximately 50% of the potential hits in the primary screen. Also, setting the hit limit further away from the mean of the population will increase the false negative rate. Therefore, this poses a dilemma as to where to set the hit limit for a Gaussian population. Setting the hit limit distant from the mean of the population will increase the activity (potency) and confidence limit of the hits (being different from the population mean) but also will increase *both* the false positive and false negative rates in a primary screen (only true for a pure Gaussian model). However, when the hit limit is set at $\sim 3\sigma_s$ or beyond, this pure Gaussian model calculation would deviate from many practical populations as those potent compounds, which are outliers of the population, should actually belong to a second population (vide infra).

Equation 3 is crucial in the evaluation of the false negative rate for any particular screen because the false negative rate, unlike the false positive rate, cannot be assessed easily in practice. Equation 3 (and its equivalent eq A5 in Appedix) provides a theoretical method for evaluating the false negative rate. However, the calculations shown in Figures 2 and 3 only represent an ideal case when the population distribution obeys a *Gaussian distribution*. In practice, various deviations from this theoretical analysis can be encountered. Several important practical considerations are discussed in the next sections.

Effects of Primary Hit Profile

The overall number of hits and the confirmation rate for these hits can be projected early in the validation process of the screen once the assay quality and the hit limit are known. Equations 1 and 3 can be employed to estimate the outcome of the screen from the validation and initial screening data. It has been indicated previously that the overall outcome from a screen depends mainly on (a) the screen assay quality, (b) the hit limit selection, and (c) the primary hit activity profile.[9] The hit profile from a primary screen is largely governed by the library diversity and concentration used in the screen. In practice, there are basically three scenarios that need to be considered.

The first scenario is when the hit limit is chosen $< 3\sigma_s$, e.g., $\sim 2\sigma_s$. This will intrinsically yield high overall hit rates ($\sim$2% for a Gaussian population distribution). This is typically true when weaker hits are desired and/or the screening quality is not superior ($0 < Z \ll 0.5$). In this scenario, the bulk population distribution (which approximates a Gaussian) dominates the calculation around the hit limit. Therefore, the confirmation rate (CR) and false negative rate (FNR) can be estimated quite well from the calculations shown in Figures 2 and 3.

In the second scenario, usually when the assay quality is very high ($Z > 0.5$) and only potent hits are of interest, the hit limit can be set at $>3\sigma_s$. In this case, the bulk of the

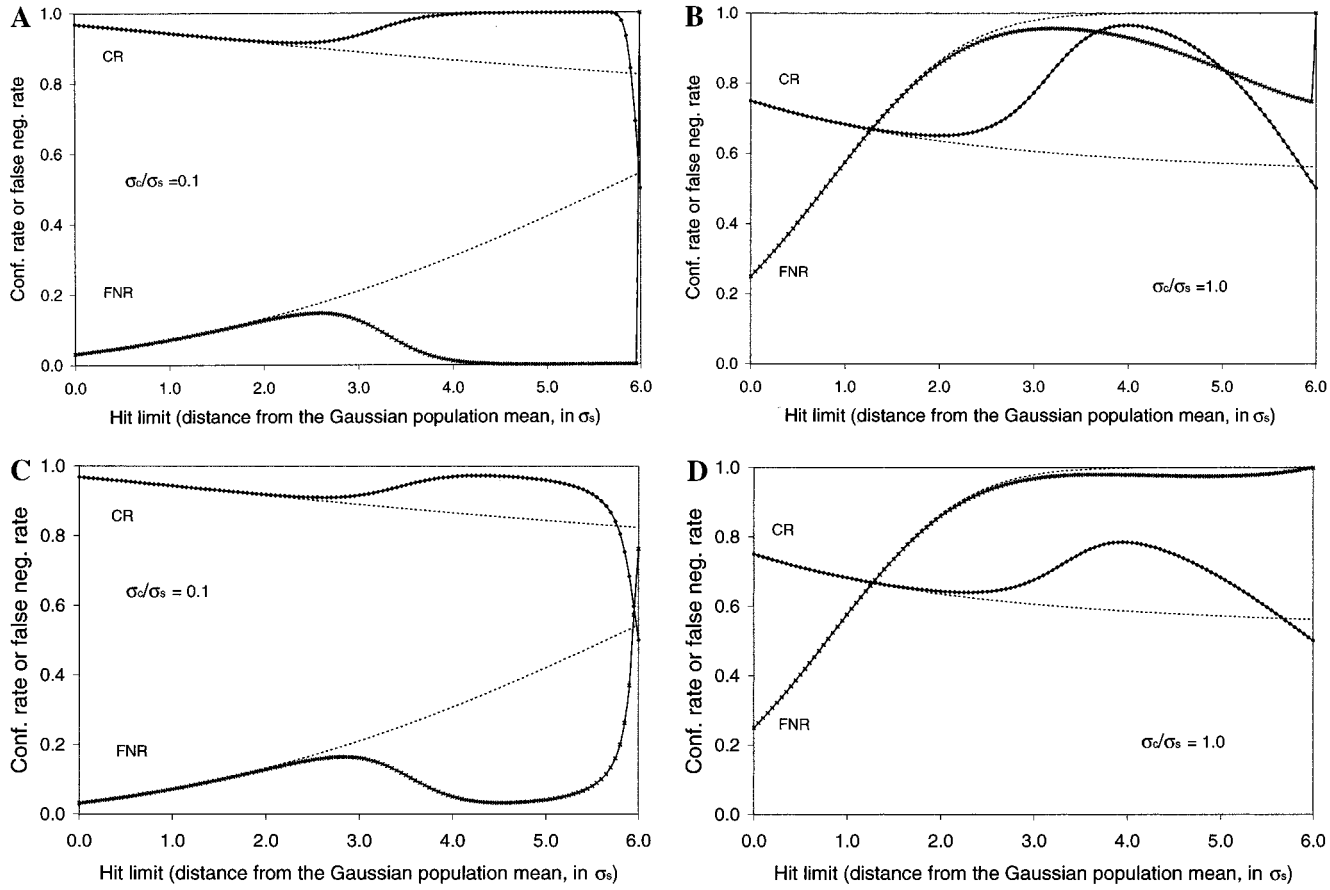

Figure 4. Calculated confirmation rate (CR, shown as ●) and the false negative rate (FNR, shown as ×) of a population when 99.9% (the bulk population) obeys a Gaussian distribution and 0.1% of the population belongs to a second distribution. A and B, the 0.1% percentile of the population resides at $6\sigma_s$ from the mean of the bulk population, as a single spike to the right of the mean. C and D, the 0.1% percentile of the population spreads *evenly* from 0 to $6\sigma_s$ to the right of the bulk population mean as a constant function. Both assay variability (σ_c/σ_s) values of 0.1 (in A and C) and 1.0 (in B and D) were used in the calculation. Dashed lines represent the CR and FNR from a pure Gaussian population, as in Figures 2 and 3. The calculation is performed using the IMSL QDAG univarate quadrature routine (see Appendix).

population contributes very little to the CR and FNR calculation since it is far away from the hit limit. Instead, the small number of potent substances near or to the right of the hit limit will contribute more to the calculation. (These potent substances usually do not belong to the bulk population distribution.) In this scenario, the CR and FNR in eqs 1 and 3 are essentially dictated by the primary hit profile, rather than the bulk of the population. Thus, the result may severely deviate from the calculations shown in Figures 2 and 3, as shown in Figure 4.

The third scenario is when the assay quality is sufficiently high ($Z \sim 0.5$) and/or the hit limit cutoff is set at $\sim 3\sigma_s$. In this case, both the bulk population and the primary hit profile affect the calculations in eqs 1 and 3. Figure 4 shows the simulated results when the bulk population (assumed to be 99.9%) obeys a Gaussian distribution at one mean (no activity) and the rest (0.1%) of the population belongs to a second distribution which is distinct from the bulk distribution (see Figure 4 caption for details). The results clearly show that the CR and FNR gradually deviate from the pure Gaussian population model when the hit limit is set away from the mean of the bulk population. Particularly when the hit limit is set at $h \sim 2.5\sigma_s$ and beyond, the hit profile (represented by the 0.1% population distribution in Figure 4) has a significant effect on the confirmation rate, false positive rate, and false negative rate. As shown in Figure 4, the confirmation rate can increase significantly (instead of decrease in a pure Gaussian distribution) as the hit limit is set further away (to some extent) from the bulk of the population. It can be seen that, while for a pure Gaussian population the confirmation rate decreases with increasing hit limit value, the confirmation rate for many populations may have a local maximum value existing at a certain hit limit value outside the major population envelope. This result agrees with the frequently encountered situations where higher confirmation rates were obtained with higher hit limit cutoff values (vide supra). It is therefore very important to set a hit limit that will most effectively identify the hits of desired potency.

Hit Confirmation Estimation Using the Primary Screening Results

After the completion of the entire primary screen, the primary hits and their activity profile become a known characteristic of the screen. From these data, the confirmation rate can be predicted based on the number of hits and their corresponding activity relative to the mean of the population and the assay quality. The formula for calculating the overall confirmation rate is reduced from eq 1 to the following discrete function

$$\mathrm{CR} = \sum_{i=1}^{n_{\max}} \frac{n_i P_i}{N_{\mathrm{tp}}} \qquad (4)$$

(equivalent to eq 1)

and the false negative rate (eq 3) becomes

$$\mathrm{FNR} = \frac{\sum_{j=1}^{n'_{\max}} n_j P_j}{\sum_{j=1}^{n'_{\max}} n_j P_j + \sum_{i=1}^{n_{\max}} n_i P_i} \qquad (5)$$

(equivalent to eq 3)

in which n_i is the number of primary hits at activity i ($\geq h$), n_j is the number of substances at activity j ($< h$), and $n_{\max}$ and $n'_{\max}$ are the maximum number of substances needed to be considered for i and j, respectively. P_i and P_j are the probability of confirmation according to the accumulative Gaussian function $P(\upsilon)$ above (Figure 1C), and N_{tp} ($= \sum n_i$) is the total number of positive hits identified from the primary screen. Similarly, the confirmation rate within a certain activity range can be evaluated when N represents the total number of hits within that range. Equations 4 and 5 are based on the specific primary screen result and therefore are more reliable and useful in post-screen hit evaluation and analysis. Importantly, eqs 4 and 5 are applicable for any population distribution as well as Gaussian.

Further Practical Considerations

In practice, various degrees of deviation from the above theoretical model for the confirmation rate can be observed. Several practical factors or limitations that need to be considered when applying the statistical model or comparing to the practical results are discussed below.

Proportional Error. When the positive and negative controls of an assay show significantly different variability, such as when there is a proportional error associated with the signal, the constant measurement error assumption no longer holds true. For example, in a specific radiometric assay, the negative control data (background) gives a low level DPM reading (e.g., 200 DPM with a standard deviation of 50 DPM), while the positive control (the full signal in the assay range) has a high level DPM reading with a higher data variation (e.g., a 5000 DPM with a standard deviation of 300 DPM). In these cases, the errors may vary linearly between those of the negative and positive control data (e.g., from 50 DPM to 300 DPM). This can obviously result in deviations from the constant error model prediction. Models with proportional error assumption are possible. However, since each different assay has its own errors spread over a specific assay dynamic range, there would be a different error "slope" for each specific assay. For this reason, a universal formula for confirmation analysis on proportional error is not trivial. In this case, a crude but simple solution for estimating the confirmation rate is to apply the constant error assumption formula with the error *at the hit limit*, which usually can be obtained by interpolating the errors of positive and negative control data.

Normality. In cases where the activity profiles of the chemical or biological library deviated from the Gaussian distribution (for example, a nonrandom, biased, or very small library), the activity profile of the population could be severely skewed or even bimodal. In these cases the real results will be significantly different from the model prediction. Best prediction results can be anticipated when the normality is good and the hit limit is set less than three standard deviations away from the population mean. However, normality has little effect on the hit identification or hit confirmation if the analysis is based on the primary screening results (i.e., using eqs 4 and 5). Furthermore, since the number of hits is usually small relative to the entire population, some degree of fluctuation from prediction is expected, even for a population with relatively good normality.

Assay Format. The choice of assay format is usually determined by the assay quality and feasibility for screening. In the above statistical model, the assay format and conditions used in primary screening and in retesting or confirmation are assumed to be the same. However, each assay format may also be biased or interfere with some set of substances in the library. Therefore, in many cases, some subset of hits, as well as the false positive or false negative hits, from a screen may be associated with the particular assay format. For instance, for a receptor binding assay, the assay can be set with either a limited amount of receptor or a limited amount of ligand, depending on the particular format used. The hits generated from these formats may not be identical. Some "undesirable" hits can also be assay format dependent. For instance, an assay format based on biotin-streptavidin interaction for a particular assay may find from a chemical library those compounds with a free biotin motif as 'hits'. (These 'hits' are really format-associated artifacts and will eventually need to be screened out in follow-up tests.) The screen process may also affect the confirmation of primary hits. For example, even within the same assay format, different pipetting methods may give different numbers of false positive hits due to reagent carry-over.

Concentration. The concentration of the collection (e.g., library compounds) at which the screening is conducted will affect the population activity profile, thus affecting the confirmation of hits. For most HTS programs, the screening is performed at one specific concentration. If only very potent hits are desired, the concentration can be set relatively low (e.g., $\sim$1 μM or lower in drug screening), when relatively weak hits are desired the concentration can be set higher (e.g., 10 μM or higher in drug screening).

Systematic Error. One of the most important issues in the HTS process is to have good control over the data variability as a function of time or reagent preparation. When applying the model for prediction or screen validation, caution must be exercised in the error assessment (i.e., standard deviations) used in these calculations. The day-to-day or batch-to-batch data variability and any possible systematic errors involved should be checked and remedied properly in order to obtain the most reliable error assessment for the screen.

Other Errors. It should also be noted that the random measurement errors of the assay and some possible systematic errors are not the only error sources encountered in

practice. Errors associated with the chemical (or biological) library used for screening (e.g., compound impurity, reactivity, degradation, interference, compound transfer error, etc.) may sometimes be the prominent sources of error. These errors (noticed or unnoticed) could give rise to seemingly abnormal hit confirmation results.

Hit Limit Selection and Assay Quality. In many practical HTS programs, the threshold for hit declaration, or the hit cutoff value, is usually a choice or compromise between practical requirements (desired activity, hit quantity or hit rate, etc.) and the assay quality considerations (the false positive and negative rates, etc.). The confirmation rate, the false positive rate, and false negative rate can be effectively optimized by reducing the data variability. This can be achieved in practice either by optimizing the assay format/conditions or by testing each individual substance multiple times. Alternatively, when the false negative hits are the major concern, a two-tier strategy can be applied. One can first set a less stringent hit limit (e.g., at $2\sigma_s$) as a filter, then confirm the primary hits by multiple tests with a more stringent hit limit (e.g., at $3\sigma_s$) in order to find most of the hits with desired activity. In this case, the false negative rate is minimized at the expense of confirmation rate.

Summary

A statistical model for evaluating the confirmation, the false positive, and false negative rates of the active substances from a high throughput screen is presented. This statistical model provides a useful tool in validation of high throughput screens as well as in post-screening data analysis. The model presented here should be generally applicable to most chemical and biological random screening systems.

Acknowledgment. The authors thank Steven J. Cottrell for performing the simulation calculations and Drs. Peter G. Schultz, Frances H. Arnold, and Anne Gershenson for reading of the manuscript and helpful discussions.

Appendix

(a) Symbols and Terms Used in the Text.

h − the hit limit or hit cutoff value in the unit of standard deviation of the population activity distribution.

u − the activity in the unit of standard deviation of the population activity distribution.

σ_c − standard deviation of the measurement for assay control data.

σ_s − standard deviation of the measurement for test population data. It contains both the assay measurement error and the variability between each member of the library (i.e., diversity or treatment effect). For a large random library, it assumes the bulk of the population distribution profile roughly approximates a Gaussian distribution.

μ_s − the mean of the population.

Z'-factor − The Z'-factor has been defined as the ratio of the separation band to the dynamic range of the assay, based on the positive control and negative control data of the assay.[9] It takes the formula

$$Z' = 1 - \frac{3\sigma_{c+} + 3\sigma_{c-}}{|\mu_{c+} - \mu_{c-}|}$$

in which μ_{c-} and μ_{c+} are denoted for the means of the negative control signal and positive control signal, respectively. The standard deviations of the signals are denoted as σ_{c-} and σ_{c+}, respectively. The Z'-factor is a simple, dimensionless, and characteristic parameter for the quality of each assay.

Z-factor − It is defined as a screening window coefficient which takes a formula similar to that of the Z'-factor.[9]

$$Z = 1 - \frac{3\sigma_s + 3\sigma_c}{|\mu_s - \mu_c|}$$

in which μ_s and μ_c are denoted for the means of the population and control (usually background) signals, respectively, and the signal standard deviations are denoted as σ_s and σ_c, respectively. This coefficient takes into account the assay signal dynamic range, the data variation associated with the sample measurement in the presence of test library, and the data variation associated with the reference control measurement.

(b) Formula Derivation for a Gaussian Population. To deduce the statistical model for the hit confirmation rate (CR), the false positive rate (FPR), and false negative rate (FNR), it is first assumed that each individual hit, when tested repeatedly, regardless of its activity, exhibits a variability profile of a Gaussian distribution, with a constant standard deviation, σ_c (the constant error assumption). Under this assumption, the probability function for hit confirmation, $P(v)$, in eqs 1 and 3 can be expressed as the cumulative function of a Gaussian distribution function for the variability of measurements, $f(w)$. It is further assumed that for any large, unbiased, random population, the activity profile from this population obeys or approximates a Gaussian distribution. Under this assumption, the formula for the population distribution, $f(u)$, can also be estimated by a Gaussian function.

Let X be the discrete variable (with a mean, μ_s) and U be the standardized discrete variables for u, the standardized variable of the Gaussian distribution of a population, $f(u)$, with a standard deviation, σ_s. Thus

$$f(u) = 1/\sqrt{2\pi}\, e^{-u^2/2} \qquad \text{(A1)}$$

where $U = (X - \mu_s)/\sigma_s$.

Also, let W and V be the discrete variables for the Gaussian and accumulative Gaussian distributions for measurement variability, $f(w)$ and $P(v)$, respectively, with a standard deviation, sc. Thus

$$f(w) = 1/\sqrt{2\pi}\, e^{-w^2/2} \qquad \text{(A2)}$$

and

$$P(v) = P(-\infty \le V \le v) = 1/\sqrt{2\pi}\int_{-\infty}^{V} e^{-w^2/2}\, dw \qquad \text{(A3)}$$

where $V = (U - h)/\sigma_c$.

Plugging these formulas in to eqs 1 and 3 in the text gives

$$\mathrm{CR} = \frac{1/\sqrt{2\pi}\int_h^\infty e^{-u^2/2}\int_{-\infty}^V e^{-w^2/2}\,dw\,du}{\int_h^\infty e^{-u^2/2}\,du} \quad \text{(A4)}$$

and

$$\mathrm{FNR} = \frac{\int_{-\infty}^h e^{-u^2/2}\int_{-\infty}^V e^{-w^2/2}\,dw\,du}{\int_{-\infty}^h e^{-u^2/2}\int_{-\infty}^V e^{-w^2/2}\,dw\,du + \int_h^\infty e^{-u^2/2}\int_{-\infty}^V e^{-w^2/2}\,dw\,du} \quad \text{(A5)}$$

Equations A4 and A5 above are derived for a screen where each substance is measured only once. If multiple independent measurements are performed for each data point, then the standard deviation of the mean of n measurements is given by $\sigma_{ave} = \sigma_c/\sqrt{n}$. Use of $V = (U - h)/\sigma_{ave}$ or $V = \sqrt{n}(U - h)/\sigma_c$ in eq A3 will yield $P(v)$ for multiple measurements.

(c) Experimental Protocol. The evaluation of the areas under the Gaussian curves was performed using either the QDAG or QDAGI routines from the IMSL Fortran library. The QDAGI routine is used to evaluate $f(u)$ in the numerator and the denominator of eq 1 and A4. The QDAGI routine uses a 21-point Gauss-Kronod rule to estimate the interval $[h, +\infty]$, after first transforming the semi-infinite interval into the finite interval [0, 1]. The function $P(v)$ was evaluated using the general-purpose QDAG routine for integrating bounded intervals. For a given hit limit and for each subinterval in $f(u)$, the integration of $P(v)$ was performed by estimating the area under the probability curve from $P(0) = 0.5$ to $P(+\infty) = 1.0$, after first dividing by the value for σ_c. QDAG is also used to evaluate $P(v)$ in the numerator and denominator of eqs 3 and A5 for the area under the probability curve from $P(-\infty) = 0$ to $P(0) = 0.5$, for each subinterval in $f(u)$ in $[-\infty, h]$. The QDAGI routine is used to evaluate $f(u)$ in eqs 3 and A5 over the intervals $[-\infty, h]$ and $[h, \infty]$.

References and Notes

(1) Dunnett, C. W. *Drug Screening: The Never-Ending Search for New and Better Drugs* in *Statistics: A guide to the Unknown*; Tanur, J. M., et al., Eds.; Holden-Day: San Francisco, 1978; pp 26−37.

(2) Devlin, J. P., Ed. *High throughput Screening: The Discovery of Bioactive Substances*; Marcel Dekker: New York, 1997.

(3) Cole, S. T.; Saint, G. I. Bacterial genomics. *FEMS Microbiol Rev.* **1994**, *14*, 139−160.

(4) Korenberg, J. R.; Chen X. N.; Adams, M. D.; Venter, J. C. Toward a cDNA map of the human genome. *Genomics* **1995**, *29*, 364−70.

(5) Fasman, K. H.; Letovsky, S. I.; Li, P.; Cottingham, R. W.; Kingsbury, D. T. The GDB Human Genome Database Anno. *Nucleic Acids Res.* **1997**, *25*, 72−81.

(6) Gordon, E. M.; Gallop, M. A.; Campbell, D.; Holmes, C.; Bermak, J.; Look, G.; Murphy, M.; Needels, M.; Jacobs, J.; Sugarman, J.; Chinn, J.; Ruhland-Fritsch, B.; Jones, D. Combinatorial organic synthesis: Application to drug discovery. *Eur. J. Med. Chem.* **1995**, *30*, S337−S348.

(7) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Application of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies and future directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(8) Oldenburg, K. R.; Kariv, I.; Zhang, J.-H.; Chung, T. D. Y.; Lin, S. *Assay Miniaturization: Developing Technologies and Assay Formats* in *Handbook of Drug Screening*; Seethala, R., Fernandes, P., Eds.; Marcel Dekker: New York, in press.

(9) Zhang, J.-H.; Chung, T. D−Y.; Oldenburg, K. R. A Simple Statistic Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screening* **1999**, *4*, 67−73.

(10) (a) Barlow, R. In *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*; Mandel, F., et al., Eds.; John Wiley & Sons: New York, 1989; pp 20−179. (b) Spiegel, M. R. In *Theory and Problems of Statistics*; McGraw-Hill Book Company: New York, 1961; pp 122−200. (c) Montgomery, D. C. In *Design and Analysis of Experiments*, 4th ed.; John Wiley and Sons: New York, 1997; pp 20−117.

(11) Sittampalam, G. S.; et al. Design of signal windows in high throughput screening assays for drug discovery. *J. Biomol. Screening* **1997**, *2*, 159−169.

(12) Burden, R. L.; Faires, J. D. In *Numerical Analysis*, 5th ed.; PWS−Kent Publishing Co.: Boston, 1993; pp 153−231.

CC9900706