Improving the Design and Analysis of High-Throughput Screening Technology Comparison Experiments Using Statistical Modeling

PHILIP W. WOODWARD,1 CHRISTINE WILLIAMS,2 ANDREAS SEWING,3 and NEIL BENSON2

Contemporary small-molecule drug discovery frequently involves the screening of large compound files as a core activity. Subsequently cost, speed, and safety become critical issues. In order to meet this need, numerous technologies have been developed to allow mix and measure approaches, facilitate miniaturization, and to increase speed and to minimize the use of potentially hazardous reagents such as radioactive materials. However, despite the on-paper advantages of these new technologies, risks can remain undefined. For example, the question of whether the novel method will facilitate identification of active chemical series in a way that is comparable with conventional methods arises. In order to address this question, we have taken the approach of carrying out experiments to directly compare the output of high-throughput screens using a given novel approach and a traditional method. The concordance between the screening methods can then be determined via comparison of the numbers and structures of the active molecules identified. This article describes the approach taken in our laboratory to minimize variability in such experiments and shows data that exemplifies the general result of lower than expected concordance. Statistical modeling was subsequently used to facilitate this interpretation. The model used β-distribution function to generate a real-activity frequency relationship with added normal random error and occasional outliers to represent assay variability. Hence, the effect of assay parameters such as the threshold, the number of real actives, and the number of outliers and the standard deviation could readily be explored. The model was found to describe the data reasonably and moreover was found to be of great utility when it came to planning further optimal experiments. A key conclusion from the model was that concordance between screening methods could appear poor even when one approach is compared with itself. This occurs simply because the result is a function of assay threshold, standard deviation and the true compound % activity. In response to this finding we have adopted alternative experimental designs that more reliably measure the concordance between screening methods. (Journal of Biomolecular Screening 2006:5-12)

Key words: high-throughput screening, concordance, statistical modeling, technology

INTRODUCTION

As the application of technology such as automated compound synthesis has advanced, it has become feasible to produce increasingly large numbers of diverse chemical compounds. As these files can be a source of novel starting points for drug discovery, screening these compounds has become part of the process in many pharmaceutical companies. However, the files can often number in the hundreds of thousands of compounds; consequently, cost, speed, and safety can become issues that must be addressed. To meet this need, numerous technologies have been developed to allow mix-and-measure approaches, facilitate miniaturization, increase speed, and minimize the use of potentially hazardous reagents such as radioactive materials. Hence, the screening scientist is often faced with a number of assay options with apparent advantages compared to existing approaches. However, data on the performance of these new approaches, in terms of identifying diverse chemical matter for drug development, are often not available. This leads to a critical question as to whether the new approach allows identification of genuinely active chemical series in a way that is comparable with conventional methods. To
address this question, an experimental approach can be taken\textsuperscript{2,3} whereby compound collections are screened using a novel approach and a comparator. The outputs in terms of active molecules identified can then be compared directly to assess the success of the novel approach relative to the conventional method. Although such experiments seem at first glance to be simple, we have found that interpretation of the results can be complex. This article shows the results of example experiments carried out in our laboratories and goes on to describe how the use of statistical modeling helped in the interpretation of the output and to design more optimal experiments. We suggest that in the search for increased productivity, statistical computer simulation and modeling can be a useful tool in the experimental design process.

**MATERIALS AND METHODS**

**Neutral endopeptidase and dry-down experiments**

Experiments were carried out as reported elsewhere.\textsuperscript{4}

**Nomenclature**

Threshold—the percent effect above which a compound well is declared a primary or secondary active.

Primary active—a compound well that is measured as over the threshold once.

Secondary active—a compound well that is measured as over the threshold twice in separate experiments.

Confirmed active—a compound that gives evidence of a dose response.

Real active—a compound that would give greater than 0% effect if all experimental errors could be removed.

Observed activity—percent effect observed.

Real activity—percent effect that would be observed if all experimental errors could be removed.

False-positive risk—the probability of a real inactive being declared a primary active due to assay variability.

False-negative risk—the probability of a real active not being declared a primary active due to assay variability.

Outlier—a percent response further from expected than can be explained by typical assay variability.

Concordance—the number of compounds identified as secondary active on all of the screening methods, expressed as a percentage of the total secondary actives.

**Threshold identification method**

To set the threshold at which a compound is declared active, we carried out experiments as follows: to 10\times 384-well plates containing randomly chosen compounds, we added either buffer control or a spike of a known active at a concentration to give approximately 67\% effect. The relevant assay components were then added, the reactions were allowed to occur, and the plates were read as appropriate. The arithmetic mean and sample standard deviation were then calculated for the spiked (M\textsubscript{67} and SD\textsubscript{67}) and nonspiked (M\textsubscript{0} and SD\textsubscript{0}) data using standard formulas.\textsuperscript{5}

The active threshold was then set according to equation (1):

\[
M_{67} - 3 \times SD_{67}.
\]

Using the mean and standard deviation of the buffer control data, M\textsubscript{0} and SD\textsubscript{0}, and assuming a normal distribution, the potential false-positive risk for this threshold can also readily be calculated using the NORMDIST function in Excel (Microsoft), which computes the integral shown in equation (2):

\[
\int_{-\infty}^{x} \left( \frac{1}{\sqrt{2\pi}} \right) e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} dx,
\]

where \(x\) = response value, \(\mu = M_0\), \(\sigma = SD_0\), and \(T\) = threshold value.

**Statistical model of screening process**

A model was developed so that a virtual comparison experiment could be simulated. Each compound screened is either inactive with a high probability (typically about 0.999), giving real activity of 0\%, or active with a complementary low probability. The real activity of simulated active compounds differs between compounds, as defined by a beta distribution\textsuperscript{6} bounded between 0 and 100, with the 1st parameter equal to 1 and the 2nd parameter fixed to represent typical results. Adding additional random error, representing assay variability, onto its real activity generated the observed activity for a specific compound. This error was represented by a normal distribution, with mean zero and standard deviation between 10 and 20. The model also allowed for the occasional outlier that causes any compound, irrespective of its true activity, to
give a response over the threshold. These were generated at random with a typical probability of 0.002.

RESULTS

To explore the performance of a novel compound dry-down procedure, we carried out a concordance experiment using neutral endopeptidase as a test system. The relative output from screening a file subset was used as a measure of success. An assay was configured for the conventional > 10-µL volume Me2SO solubilized method and another one for a miniaturized microliter scale method employing compound dry down (see Materials and Methods), to give 2 assays that were then used to screen the same set of compounds. To minimize variables, the comparison screens were configured as far as possible to facilitate equal probability of identifying actives of a particular potency independent of differential assay performance. Specifically, the biochemical kinetics was balanced, and the threshold for each screen was set using the approach described in the Materials and Methods section. The intention was that this experimental design should ensure, as far as possible, that the vast majority of sub 5-µM actives would be identified in each screen. Typical data are shown in Figure 1.

Using the thresholds derived experimentally, we then carried out a screen of several hundred thousand compounds with 1 compound per well for both targets. The screens were judged to be of good quality, as shown by the Z' factor calculated from controls across the screens; the Z' factor was generally > 0.4 (Fig. 2). All primary active compound wells were then restested to generate a list of secondary actives (see Fig. 3 for a summary of the workflow). These were then compared to establish the concordance. The results showed that the concordance was around 30% (Fig. 4). One interpretation of these data is that the novel approach yields a result inconsistent with expectation. Potentially, this could arise due to false positives and negatives occurring as a result of the specifics of the methodology. However, considering first that there was a tendency for the alternative approach to identify similar numbers of actives and, second, because of the typical potency of most compounds observed in our high-throughput screens (cf. Fig. 5), we postulated that some of the apparent lack of overlap could be due in large part to the natural variability of the assay. To explore this further, we have developed a statistical model of concordance experiments as described in the Materials and Methods section. Compounds were classified as either inactive or active. The inactive compounds give nonzero observed activities due to assay noise. In terms of the active pool, our experience and intuition indicate that active compounds are not equally potent; indeed, a continuous distribution of activity is most often observed, with the frequency increasing as potency decreases (extreme outliers have been removed as these are not believed to represent true activity). A beta distribution can be used to represent such a distribution of true activity. Figure 6 illustrates the general shape of this distribution when the 1st parameter is fixed at 1. However, measured activity varies not only due to the real differences in potency of each com-
pound but also due to run-to-run measurement variation. Unpublished internal studies have shown that this assay variability can be approximated by a normal distribution with standard deviation between 10% and 20% for most screens. Note that this degree of variation is not based on replicated controls on the same plate but rather on replicated compounds assayed on different plates and different days. Using the beta distribution and accounting for these normal distributions, we have found that this model can be used to represent typical screening data using 2.7 as the 2nd parameter in the distribution (cf. Fig. 5 vs. Fig. 7).

An additional component of the model needs to account for each screen having a different inherent hit rate due to the relative number of compounds that are truly active against that target. This is modeled by specifying the proportion of the compounds that are treated as being active and hence having a true activity response greater than 0%. Our experience showed that the hit rate typically lies between 0.01% and 1%. Since these percentages represent the relative number of compounds that repeatedly pass the activity threshold, it is necessary to adjust these values upward in our model because many of the active compounds will be too weak to meet this criterion. That is, referring to Figure 6, it is seen that many of the real actives will have real activity relatively close to 0 and hence will not be detected as secondary actives. Hence, our current perception of the typical number of active compounds in a library will be less than the number of real actives as defined here. The necessary adjustment is easily calculated once the beta distribution that represents the distribution of activity and the threshold are defined.

Due to the large number of compounds that are tested in a screen, outliers are expected, so this feature is also modeled. Outli-
ers (i.e., compound wells that give a response above the threshold irrespective of its potency) are simulated with a low probability (a typical value consistent with our experience is 0.2%).

A threshold of 40% was used for the majority of the simulation studies, and the effects of changing the assay variation (SD; 10%, 15%, and 20%), the outlier rate (0%, 0.2%, and 0.5%), the screen hit rate (0.01%-1%), and the number of screening methods being compared (2 or 3) were included as parameters in the model. The effect of changing the threshold within our view of realistic limits (30%-50%) was also assessed.

The model showed that the dominant factor, over the ranges studied, is the assay variation (cf. Fig. 8A,B). This was exacerbated when the true hit rate was low. It was also clear from these simulations that, even when comparing an approach with itself, we should not expect to see a very high concordance percentage. A typical study comparing 2 identically performing approaches, with a run-to-run standard deviation of 15%, should not be expected to give greater than 50% concordance, decreasing to about 33% when 3 screening methods are compared. The simulations also show that using screens that exhibit low genuine active rates will tend to yield lower concordance due to the fact that false positives make up a more significant proportion of the secondary actives in these cases. Changing the threshold (cf. Fig. 9) did not improve the concordance. For screens rich in actives (about 1%), the choice of threshold had little effect on the concordance. But as the percentage of real actives decreased, a lower threshold dramatically re-

---

**FIG. 8.** (A) Graph showing effects of run-to-run assay variation (SD), outliers, and hit rate of the screen on the concordance percentage that can be expected when 3 identical screening methods are studied. (B) Graph showing effects of run-to-run assay variation (SD), outliers, and hit rate of the screen on the concordance percentage that can be expected when 2 identical screening methods are studied.
duced the expected concordance. This latter result is due to the additional false positives that are being generated by the lower threshold. In our example concordance experiment (Fig. 4A), the 30% overlap observed can now be evaluated in the context of the model. The SD within the sample areas was of the order of 15%, and an outlier rate of 0.2% was estimated from legacy Pfizer data sets for random compounds replicated on different days. Based on existing knowledge, the expectation for this particular target was a high hit rate of the order of 0.3%, leading to an expected concordance in the 40% to 50% region (Fig. 8B), assuming there were no artifactual positives in the actives identified or nonnormal variability and that the variability and thresholds in the 2 screens were similar.

**DISCUSSION**

In the development of high-throughput screening as a drug discovery tool, a plethora of new technologies and approaches have come to the fore. All can offer apparent advantages over one another or relative to more traditional methods. However, the critical question of quantifying how good this technology will be in terms of identifying all genuinely active compounds often remains unspecified. Careful consideration of this apparently simple question leads to the need for an experiment that can give data on the performance in terms of identifying active compounds with a range of physical properties in respective formats. In addition, we believe that as structure-activity relationships and ligand efficiency become data of interest, the experiments need also to quantify the sensitivity of respective technologies to less potent compounds.

To address these questions for a particular technology of interest, we have taken the approach of screening a collection of compounds against a given target using the new technology and comparing this with a more traditional approach. In the case discussed in this article, the approach of interest was the use of dry-down procedures in single microliter assay scale assays. The experiments were designed to capture any compound property-specific effects by screening as wide a range of chemotypes as possible and to reveal the relative success of identifying the weaker actives.

Accurately comparing approaches is dependent on the minimization of differences between the experiments. To this end, we set up the biochemistry and sensitivity to active compounds to be as similar as possible for the approaches compared. Firstly, the biochemical kinetics were balanced; that is, S/K_M ratios were identical and buffer compositions were kept as similar as possible. Secondly, as different assays will exhibit different performance statistics and therefore potentially dissimilar sensitivities to active molecules, we used an experimental approach to independently define the threshold needed for each screen. The experiment involved generating 2 populations of data, one at 0% effect and the other at 67% effect, as described and shown in Figure 1: 67% was chosen as this represents the approximate theoretical percent effect of a 5-µM active at 10 µM (the dose used in our screens), calculated from a standard 4-parameter logistic function. IC_{50} was assumed to approximate K_M and slope to be unity. The 5-µM limit was derived from the hypothesis that compounds with equal to or less than this affinity represent optimal starting points for further development, and hence our screens are designed to maximize the probability that these are identified. Random compounds were added to both the 0% and 67% effect test wells as an attempt to capture errors due to compound addition, biophysical interferences, and any other source of variability. From the data and equation (1), a threshold may then be calculated that ensures as far as possible that for each screen, the vast majority of 5-µM actives will be identified. In addition, the data allow an estimate of the likely false-positive risk associated with the threshold chosen (see equation (2)). Typically, we have found that our assays have median Z' factors of

![Graph showing effects of the threshold and hit rate of the screen on the concordance percentage that can be expected when 2 identical screening methods are studied (SD = 15%, outliers = 0.2%).]
the order of 0.6 to 0.7, and false positives arising purely from this approach are low. If the situation did arise that the threshold set was likely to give rise to significant numbers of false positives, the experiment described allows a decision to be taken on a quantitative basis. For example, the cost of screening the number of false positives predicted could be weighed against the cost of working to improve the assay versus a compromise on the ability to detect all sub 5-μM actives.

Given the experimental design and the data quality, our results were not as we had anticipated (cf. Fig. 4) in that data concordance of the order of 30% was observed. One possible conclusion was that the novel approach yielded lower numbers of true actives and was producing considerable numbers of false positives. However, before invoking this explanation, we considered an alternative additional hypothesis that these results could be rationalized in large part via the natural variability of the assays. Specifically, we noted that the control experiment of comparing an approach versus itself is not typically carried out, and hence we did not have a measure of the effect of random noise on any outcome. It might be expected that in this case, something close to 100% concordance should be observed. However, given that each screen has an intrinsic standard deviation of the order of 10% to 20%, and because our experience has shown that the percent effect of actives varies with a bias toward the weaker potencies, we postulated that the actual observed concordance could be a function of assay variability and the percent effect of the compounds relative to the threshold. One solution to this would be to restrict analysis to those compounds that give percent effect in the higher potency range. However, in our view, this confounds the experimental strategy in the sense that the compounds in the 40% to 60% range are of interest and therefore so is the performance of the new approach in this area. In addition, as the percent effect approaches 100%, the relationship between potency and effect changes, leading to difficulties in drawing conclusions. Another alternative is to obtain a baseline measure of concordance by comparing a particular approach with itself—that is, to run the whole screening process through the same approach 2 or 3 times and analyze as if the data had come from different screening methods. Clearly, however, this would be a time-consuming and expensive exercise; thus, as an alternative, we have used a statistical model of the process, as described in the Materials and Methods section.

To produce a working model for hypothesis testing, we have made certain assumptions. Most critically, the shape of the distribution of real activity (cf. Fig 6) is, to some extent, speculation because to our knowledge there is no underlying theory that describes this. However, an analysis of a range of our screens has shown behavior consistent with this (cf. Fig. 5), as have the analyses of others. In addition, we have found practically that the model can explain the gross features of the data we have observed to date and helps us to understand the complexities of the experiments. It would be possible to alter the parameters of the beta distribution to represent other activity distributions. Certain questions do, however, remain difficult to address; in particular, the potential contribution of artifacts can complicate analysis. Hence, the model cannot explain all of the observed variability, and the observed and predicted results differed by ca. 10% to 20% (a 30% observed concordance vs. a predicted 40%-50%). However, our view is that the model at least sets the expectation of an upper limit for concordance that was intuitively less than expected. That the observed concordance was less than predicted by the model is to be expected because of the known contribution of artifacts and nonnormal variability. Ultimately, a key benefit of the model was to lead us to the conclusion that the 30% concordance was, potentially at least, consistent with the novel approach performing comparably with the conventional approach. Subsequent experiments were consistent with this conclusion. In addition to facilitating this understanding, the model has also allowed for a range of conditions to be readily explored that have led to the design of more efficient experiments for any future work. For example, we have concluded that screens should be used that will meet the dual need for a relatively high number of actives in the screened population (> 0.1%) and that also display activities spanning the range of percent effects of interest (e.g., 40%-100%). Instead of converting the continuous percent effect values into binary hit/miss results, as is done in the concordance analysis described above, the concordance between the 2 approaches should be assessed using traditional statistical measures such as x-y scatter plots, regression, and correlation coefficients. Also, we have found that experiments comparing the IC_{50} of confirmed actives for a range of compound types are a more powerful way of drawing clear conclusions than using the activity response obtained from single-point high-throughput assays.

ACKNOWLEDGMENTS

We thank Simon Dales (Pfizer Sandwich Discovery Biology) and Graeme MacDonald (Medicinal Technologies, Pfizer, Sandwich) for providing the NEP enzyme. We also thank Gareth Williams, Malcolm Wickenden, and Philip Gibbon (Lead Discovery Technologies, Pfizer; Sandwich) for preparing and screening the plates and express our gratitude to Richard Lyons (Non-Clinical Statistics Group) for his valuable help and advice with the beta distribution model.

REFERENCES