Statistical Analysis of Systematic Errors in High-Throughput Screening

DMYTRO KEVORKOV[1] and VLADIMIR MAKARENKOV[2]

High-throughput screening (HTS) is an efficient technology for drug discovery. It allows for screening of more than 100,000 compounds a day per screen and requires effective procedures for quality control. The authors have developed a method for evaluating a background surface of an HTS assay; it can be used to correct raw HTS data. This correction is necessary to take into account systematic errors that may affect the procedure of hit selection. The described method allows one to analyze experimental HTS data and determine trends and local fluctuations of the corresponding background surfaces. For an assay with a large number of plates, the deviations of the background surface from a plane are caused by systematic errors. Their influence can be minimized by the subtraction of the systematic background from the raw data. Two experimental HTS assays from the ChemBank database are examined in this article. The systematic error present in these data was estimated and removed from them. It enabled the authors to correct the hit selection procedure for both assays. (*Journal of Biomolecular Screening* 2005:557-567)

Key words: high-throughput screening, systematic error, background evaluation, trend-surface analysis

INTRODUCTION

HIGH-THROUGHPUT SCREENING (HTS) is an effective drug discovery method. This technology allows for screening of more than 100,000 compounds a day per screen. A typical HTS operation in the pharmaceutical industry generates approximately 50 million data points per year.[1] Such a great amount of experimental data requires an efficient automatic routine for the selection of hits. Unfortunately, random and systematic errors can cause a misinterpretation of HTS signals.

Various methods for quality control and correction of HTS data have been proposed in the scientific literature. Their descriptions can be found in the articles by Heuer et al.,[1] Gunter et al.,[2] Brideau et al.,[3] Heyse,[4] and Zhang et al.[5,6] It is essential to identify and compensate for the systematic variability of HTS measurements. Heuer et al.[1] and Brideau et al.[3] showed examples of systematic trends in HTS plates; the trends of this kind are present in all plates of an assay. The systematic errors caused by aging, reagent evaporation, or decay of cells can be recognized as smooth trends in plate means/medians over a screen. Errors in liquid handling and malfunction of pipettes can also generate localized deviation of expected data values. Variation in incubation time, time drift in measuring different wells or different plates, and reader effects may be recognized as smooth attenuation of measurement over an assay.[1] An example discussed by Brideau et al.[3] illustrates a systematic error caused by the positional effect of the detector. Throughout the entire screening campaign involving more than 1000 plates, signal values in row A were on average 14% lower than those in row P (see Brideau et al.[3], Figure 1). Such effects may have a significant influence on the hit selection process. They can cause either underestimation (false negatives) or overestimation (false positives) of measured data.

The aim of this work was to develop a method for minimizing the impact of systematic errors in the analysis of HTS data. A systematic error can be defined as a systematic variability of values among all plates of an assay. We will show how a systematic error can be detected and its effect removed by analyzing the background patterns in plates of the same assay.

MATERIALS AND METHODS

Experimental data

We have selected for evaluation freely available HTS data from the collection of the ChemBank database (http://chembank.med.harvard.edu/). This data bank has been maintained by the Institute for Chemistry and Cell Biology at Harvard Medical School (ChemBank Development Team, Institute for Chemistry and Cell Biology, Harvard Medical School, 250 Longwood Ave, SGM 607,

[1]Laboratoire LACIM and [2]Département d'informatique, Université du Québec à Montréal, Canada.

Received Nov 24, 2004, and in revised form Feb 23, 2005. Accepted for publication Mar 21, 2005.

Journal of Biomolecular Screening 10(6); 2005
DOI: 10.1177/1087057105276989

Boston, MA 02115). It contains public HTS data from various high-throughput chemical screens. Among available HTS assays, we have chosen for our analysis the 2 largest data sets.

The 1st one is provided by the Chemistry Department of Princeton University (http://chembank.med.harvard.edu/screens/screen_table.html?screen_id=76). It consists of a screen of compounds that inhibit the glycosyltransferase MurG function of *Escherichia coli*. The experimental data for 164 plates were considered in this study. According to the ChemBank description, this assay has been obtained during the screen that has measured the binding of MurG to a fluorescent (fluorescein-labeled) analogue of UDP-GlcNAc. Screening positives are the compounds that inhibit binding of GlcNAc to MurG. The data set contains the measurements for the tested compounds and the hits selected by the screen authors. No information about controls and the hit selection procedure is provided for these data, and there is no reference for the article describing the experimental details given on the ChemBank Web site. After the literature search, we have found the reference[7] discussing these experimental data.

The 2nd assay considered was provided by Dr. Robert Shapiro from Harvard Medical School (http://chembank.med.harvard.edu/screens/screen_table.html?screen_id =59). It consists of 59 plates. It is a primary screen for the compounds that inhibit the activity of human angiogenin, a protein with RNase activity that can induce angiogenesis. Similar to the previous assay, this data set contains measurements for the tested compounds and the hits identified; no further information about this assay is given. We have found the reference[8] discussing these experimental data.

Experimental procedure

To detect a systematic error, the following assumptions about HTS data have been made:

1. screened samples can be divided into active and inactive,
2. most of the screened samples are inactive,
3. values of the active samples differ substantially from the inactive ones,
4. samples are arranged in a 2-dimensional format and are operated in sequence,
5. systematic error causes a repeatable influence on the measurements in all plates, and
6. samples are located randomly within plates.

The first 3 assumptions divide the samples into 2 groups. The bigger group contains inactive samples. Their values, measured for a large number of plates, are similar, and their variability is caused mainly by systematic errors. Therefore, inactive samples can be used for the computation of the background. In the ideal case, the background surface is a plane. Random errors produce residuals that compensate each other during the computation of the mean background for a large number of plates. Systematic errors generate repeatable local artifacts and smooth global drifts, which become more noticeable while computing a mean background.

The analysis of experimental HTS data requires a preprocessing to ensure the statistical meaningfulness and accuracy of the background analysis and the hit selection. If the experimental HTS data have a Gaussian distribution, a logarithmic transformation can be applied prior to normalization. This transformation makes normalized data additive and renders variation more independent of absolute magnitude. The following main steps were performed in this study:

- normalization of experimental HTS data,
- outlier elimination,
- topological analysis of the background,
- elimination of systematic errors,
- selection of hits, and
- analysis of hit distribution.

Normalization

Plate means and standard deviations vary from plate to plate. Therefore, to compare and analyze together experimental data from various plates and, consequently, to generate a statistically correct background, all measurements should be normalized. We have considered and compared 2 normalization procedures.

Mean centering and unit variance standardization. This normalization procedure, also known as "normalization to zero mean and unit standard deviation" or "zero mean and unit variance standardization," ensures that all processed elements are transformed in such a way that the mean value of the normalized elements is zero, whereas the standard deviation and the variance are equal to unity. The mean value μ of n elements x_i can be computed as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

and the standard deviation σ is as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}. \tag{2}$$

Applying the following formula, we can normalize the elements of the input data:

$$x'_i = \frac{x_i - \mu}{\sigma}, \tag{3}$$

where x_i is the input element, x'_i is the normalized output element, μ is the mean value, σ is the standard deviation, and n is the total number of elements in the plate. The output data conditions will be $\mu_{x'} = 0$ and $\sigma_{x'} = 1$.

Normalization to interval. This normalization procedure determines the maximum x_{max} and minimum x_{min} values and ensures that all input elements are transformed proportionally into predefined upper and lower limits. It performs a transformation of the input elements x_i into the normalized output elements x'_i:

$$x'_i = \frac{(x_i - x_{\min})(L_{\max} - L_{\min})}{x_{\max} - x_{\min}} + L_{\min}, \tag{4}$$

where x_i is the input element, x'_i is the normalized output element, L_{max} and L_{min} are the predefined upper and lower limits, and x_{max} and x_{min} are the maximum and minimum values of the input elements. In our experiments, L_{max} and L_{min} were set to 1 and –1, respectively.

The comparison of both methods has not demonstrated any significant difference between them. In our study, the computed backgrounds obtained with the 2 normalization procedures had the similar shapes. However, the application of mean centering and unit variance standardization would be more convenient for the background generation. The main advantage on this approach is that all mean values of the normalized plates' measurements equal zero. Therefore, the mean value of the overall evaluated background will be also zero. This gives the possibility for an accurate data correction by the direct subtraction of the evaluated background from the normalized experimental data. Another advantage of mean centering and unit variance standardization is that the impact of outliers on the normalized data would be much lower compared to normalization to interval.

Background evaluation

The computation and topological analysis of the background is essential for the development of a method that detects local background effects and automatically compensates for systematic errors. We define an assay background as a mean of the normalized plate measurements:

$$z_i = \frac{1}{N}\sum_{j=1}^{N} x'_{i,j}, \tag{5}$$

where $x'_{i,j}$ is the normalized value at well i of plate j, z_i is the background value at well i, and N is the total number of plates. The bigger the number of plates, the more meaningful the background formula is. For a large number of homogeneous plates (>100), a sufficient number of low values not corresponding to hits will compensate high values of hits.

If an assay contains a small number of plates (<100), high values of hits and outliers can have a negative influence on the background surface and create false patterns not existing in the reality. In such cases, we propose to eliminate them from the background analysis. In our study, we have considered 3 cases of hit and outlier elimination. Thus, we can cut the hit and outlier values that exceed either 1σ or 2σ, or 3σ deviations from the mean of each plate. For instance, in the case of 3σ elimination, formula 5 can be rewritten as follows:

$$z_i = \frac{1}{N - N_i^{h/o}} \sum_{\substack{j=1,\\ i \neq hit\,/\,outlier}}^{N} x'_{i,j}\ , \tag{6}$$

where $x'_{i,j}$ is the normalized value at well i of plate j after the elimination of hits and outliers defined as (hit/outlier $\in [-\infty; (\mu - 3^*\sigma)] \cup [(\mu + 3^*\sigma); +\infty\,]$), z_i is the background value at well i, N is the total number of plates, and $N_{\mathrm{i}}^{\mathrm{h/o}}$ is the total number of hits and outliers at well i over N plates.

Topological analysis of the background surface

To discover general trends and local effects characterizing the evaluated background, we carried out its trend-surface analysis.[9,10] This global surface-fitting procedure is widely used in geographical applications, digital surface modeling, and biostatistics. A polynomial function is usually chosen for the approximation of experimental data. The polynomial can be expanded to any desired degree, and the unknown coefficients can be found by the polynomial least squares fit or by the multiple linear regression. The following formula presents the polynomial function of the 5th degree for a 2-dimensional surface:

$$\begin{aligned}
&Z(X,Y) = a_0 + a_1X + a_2Y + && \text{1st degree}\\
&a_3X^2 + a_4XY + a_5Y^2 + && \text{2nd degree}\\
&a_6X^3 + a_7X^2Y + a_8XY^2 + a_9Y^3 + && \text{3rd degree}\\
&a_{10}X^4 + a_{11}X^3Y + a_{12}X^2Y^2 + a_{13}XY^3 + a_{14}Y^4 + && \text{4th degree}\\
&a_{10}X^5 + a_{11}X^4Y + a_{12}X^3Y^2 + a_{12}X^2Y^3 + a_{13}XY^4 + a_{14}Y^5 && \text{5th degree.}
\end{aligned} \tag{7}$$

In our analysis, we defined X as a column number, Y as a row number, and Z(X,Y) as a measured value at the well located in the intersection of column X and row Y.

To find the polynomial coefficients and to validate the correctness of our procedure, we used both the polynomial least squares and the multiple linear regression methods. The polynomial least squares proceeds by solving a set of linear equations that includes the sums of powers and cross-products of the values of X, Y, and Z. As an example, we examine the computation of the coefficients for the 2nd-degree polynomial using the polynomial least squares. We have to solve the system of 6 linear equations presented in a matrix format (see Exhibit 1) where n is the number of wells in each plate. The solution of this system provides the best fit according to the least squares criterion. As a result, we obtain 6 optimal least square coefficients (a_0, a_1, a_2, a_3, a_4, a_5) for the 2nd-order polynomial presented in formula 8:

$$Z(X,Y) = a_0 + a_1X + a_2Y + a_3X^2 + a_4XY + a_5Y^2. \tag{8}$$

It is worth noting that both, the polynomial least squares and the multiple linear regression, have produced the identical results expected. Here, the R^2 and F statistics were used to determine the most appropriate polynomial degree.

Exhibit 1

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} Y_i^2 \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} X_i^3 & \sum_{i=1}^{n} X_i^2 Y_i & \sum_{i=1}^{n} X_i Y_i^2 \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} Y_i^2 & \sum_{i=1}^{n} X_i^2 Y_i & \sum_{i=1}^{n} X_i Y_i^2 & \sum_{i=1}^{n} Y_i^3 \\ \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i^3 & \sum_{i=1}^{n} X_i^2 Y_i & \sum_{i=1}^{n} X_i^4 & \sum_{i=1}^{n} X_i^3 Y_i & \sum_{i=1}^{n} X_i^2 Y_i^2 \\ \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} X_i^2 Y_i & \sum_{i=1}^{n} X_i Y_i^2 & \sum_{i=1}^{n} X_i^3 Y_i & \sum_{i=1}^{n} X_i^2 Y_i^2 & \sum_{i=1}^{n} X_i Y_i^3 \\ \sum_{i=1}^{n} Y_i^2 & \sum_{i=1}^{n} X_i Y_i^2 & \sum_{i=1}^{n} Y_i^3 & \sum_{i=1}^{n} X_i^2 Y_i^2 & \sum_{i=1}^{n} X_i Y_i^3 & \sum_{i=1}^{n} Y_i^4 \end{bmatrix} * \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Z_i(x,y) \\ \sum_{i=1}^{n} X_i Z_i(x,y) \\ \sum_{i=1}^{n} Y_i Z_i(x,y) \\ \sum_{i=1}^{n} X_i^2 Z_i(x,y) \\ \sum_{i=1}^{n} X_i Y_i Z_i(x,y) \\ \sum_{i=1}^{n} Y_i^2 Z_i(x,y) \end{bmatrix}$$

Elimination of systematic error and analysis of hit distribution

The presence of systematic error can be detected during the analysis of the hit distribution. If systematic error is absent, average numbers of selected hits for each well should be similar when a large number of homogeneous plates (>100) are considered. A big variation of hit numbers indicates the presence of systematic error. We examined the hit distribution by rows and columns as well as analyzed the number of hits at each well to prove the presence of systematic errors in experimental assays.

As we mentioned above, deviations of the background surface from a zero plane reflect the influence of systematic errors on experimental measurements. Therefore, we can correct the raw HTS data by subtracting the evaluated background from the normalized values of each plate of an assay. Then, we can reassess the background surface and the hit distribution obtained after the application of this correction procedure.

The evaluation of the hit distributions in the raw and corrected data sets is a 2-fold process. The subtraction of the evaluated background changes the standard deviation in the data set under consideration. The residuals, and hence their distribution, of the background-corrected data set will be changed. Generally, new standard deviations, that are calculated from the trend-corrected data, will be smaller. Obviously, the computation of the standard deviations and the hit selection are done separately for the original and corrected data sets.

RESULTS

Assay 1

In Figure 1, we present the evaluated background surface and its approximation by the 4th-degree polynomial computed for the assay comprising the data for compounds that inhibit the glycosyltransferase MurG function of *E. coli*. To estimate this background, we used normalized measurements for 164 plates containing 16 rows and 22 columns each. The total number of plates for this assay was 208, but 44 of them were excluded from our analysis because of the presence of missing rows or columns. Formula 5 was applied to compute the evaluated background. No hit/outlier elimination was done in this case.

As mentioned above, we have considered 3 cases of hit elimination. The background surface computed for 1σ elimination is presented in Figure 2a (the values that deviated for more than 1σ from the plate means were removed from the background computation). The solution surface looks like a flat surface with some minor local effects on the edges. In contrast to the background surface for 1σ hit elimination, the evaluated backgrounds for 2σ and 3σ hit elimination (see Figs. 2b and 2c) as well as the background surface computed with no hit elimination at all (see Fig. 2d) demonstrate the substantial deviations from a plane (i.e., they depict the presence of an important systematic error). Consequently, we have assumed that the 1σ elimination procedure removes not only the hit impact but also the impact that is due to systematic error. Therefore, the background determined using 1σ hit elimination was not suitable for the evaluation of systematic errors present in this assay.

We have to mention that Figure 2d corresponds to Figure 1b. Figure 1b has the background variation axis inversed; this presentation enabled us to compare the background surface to the hit distribution surface (Fig. 4b). The ChemBank Web site contains no information on how the hits were selected by the researchers who conducted the experiments. After the analysis of the indicated hits, we supposed that they selected as hits the values that deviated for more than 1σ from the plate means. The procedure using 1σ hit selection gave us the results similar to those presented by the authors of this assay. We carried out the analysis of the hit distribution by columns and rows, selecting as hits the values deviating for more

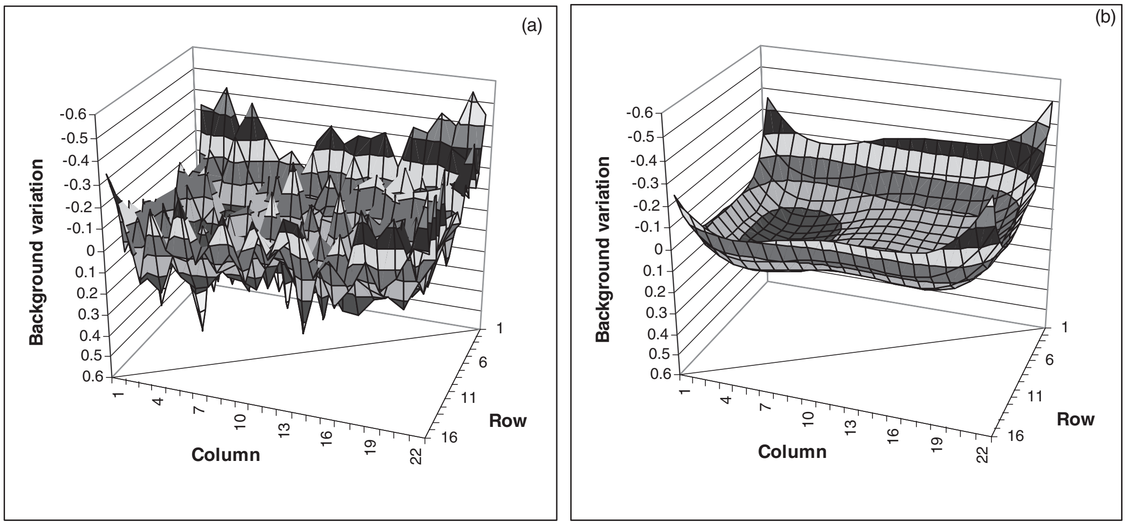

FIG. 1. (a) Evaluated background surface of assay 1 (164 plates) and (b) its approximation by the 4th-degree polynomial.

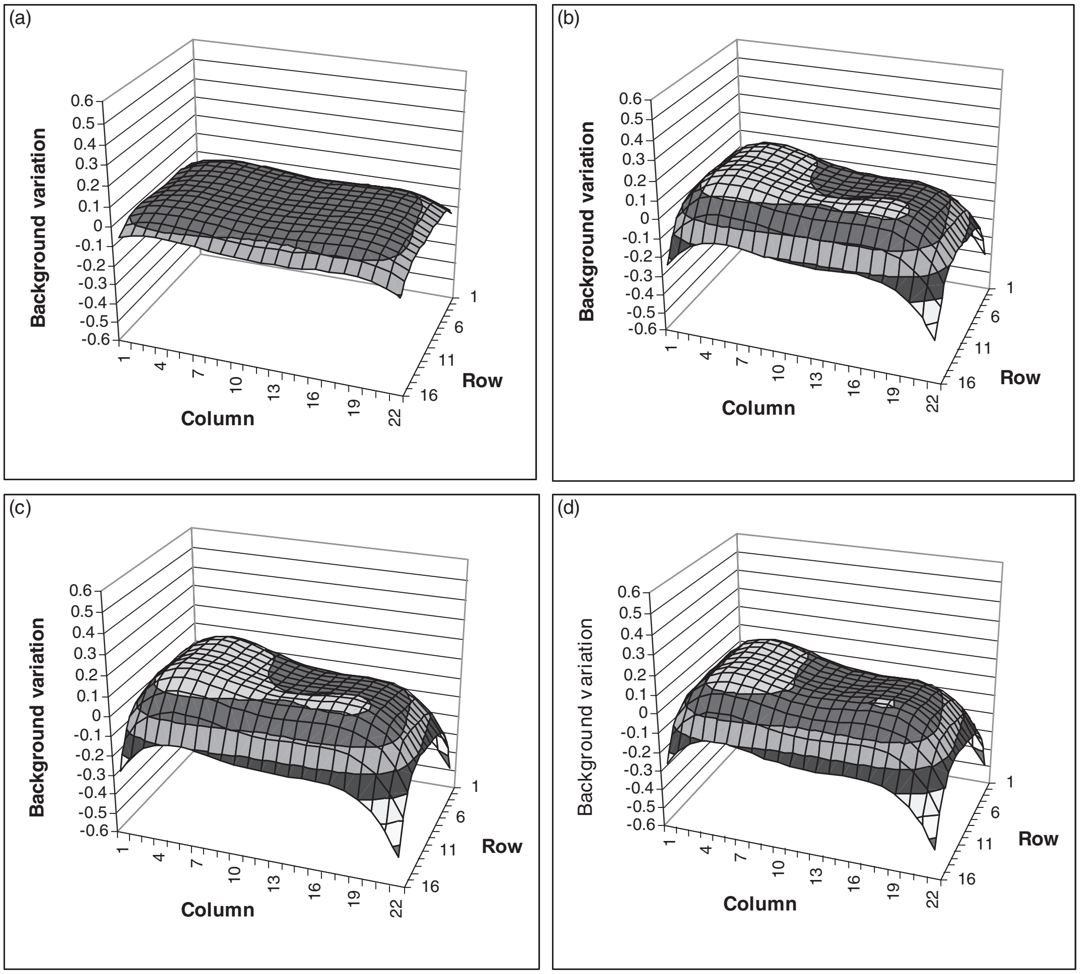

FIG. 2. Trend-surface approximation of the computed background of assay 1 (164 plates) with (a) 1σ, (b) 2σ, (c) 3σ hit elimination, and (d) with no hit elimination at all.

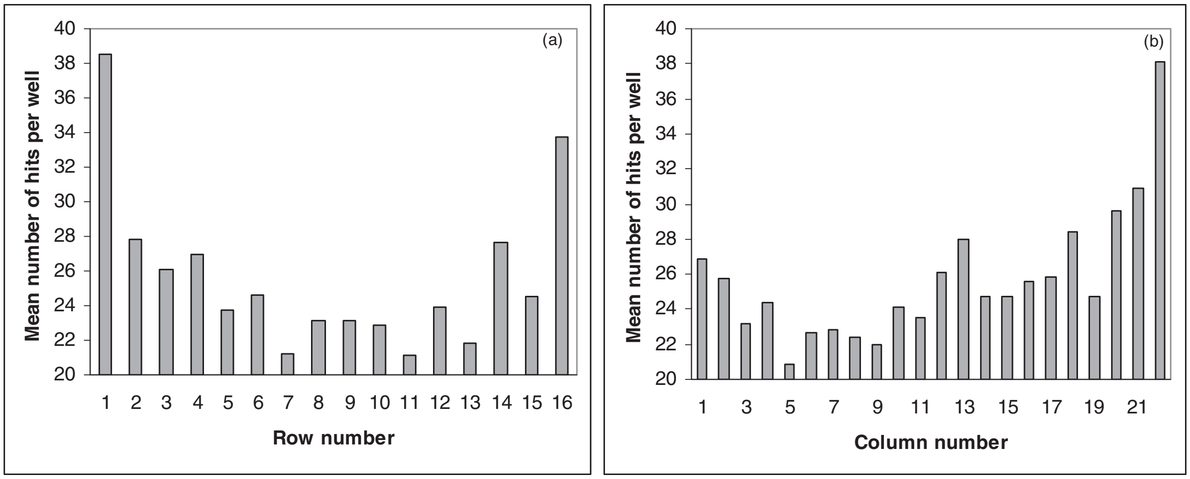

FIG. 3. Hit distribution by (a) rows and (b) columns of assay 1 computed for 164 plates.

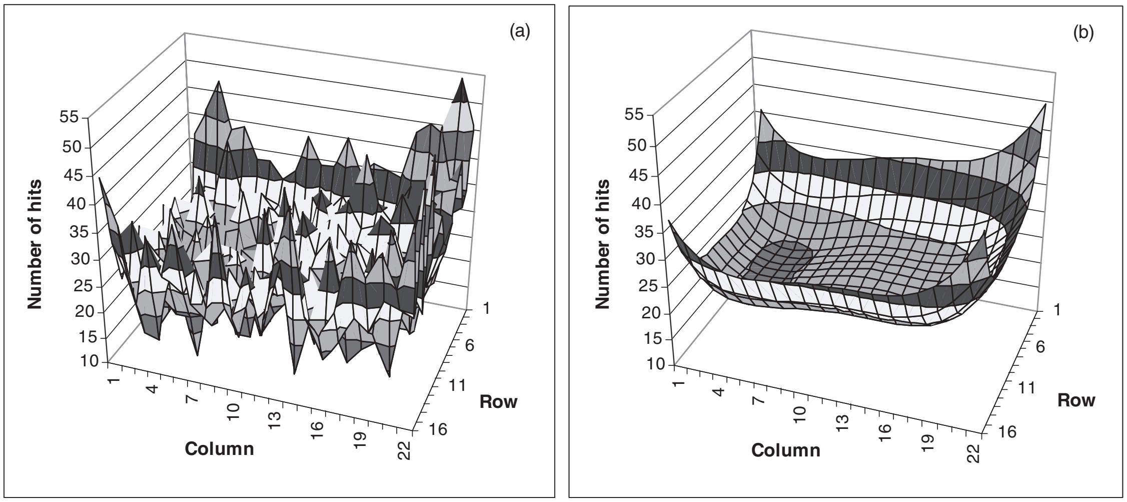

FIG. 4. (a) Hit distribution surface of assay 1 and (b) its approximation by the 4th-degree polynomial.

than 1σ from the plate means. This distribution (presented in Fig. 3a) clearly demonstrates the difference between the number of hits on the edges and in the middle of the plates (e.g., the average number of hits in row 1, computed over 164 plates, was 38.5; in contrast, the average numbers of hits in rows 7 and 11 were 21.2 and 21.1, respectively). Such a difference is unlikely due to random errors and, in our opinion, is caused by a systematic error of the measurements. To conduct a more detailed analysis, we determined the number of hits at each well. The resulting surface is presented in Figure 4a. To detect the tendencies of the hit distribution, we also carried out the trend-surface analysis of this pattern. The hit distribution surface fitted by the 4th-degree polynomial is shown in Figure 4b. It is easy to see that the trends of the hit distribution surface shown in Figure 4b correlate with the trends of the background surface illustrated in Figure 1b.

To remove the effect of systematic errors, we subtracted the overall evaluated background (without hit/outlier elimination) from each of the 164 plates used in our study. The corrected values were used for further analysis. Using the above-described procedure, new hits were selected, and their distribution was reexamined. In Figure 5, we compare the distribution of hits by rows and columns for the raw and corrected data. The hit distribution for raw data, plotted as columns, corresponds to that presented in Figure 3b. The data corrected by the subtraction of the evaluated background are depicted by a solid line with squares. We also evaluated the data correction made by the subtraction of the approximated background. The obtained hit distribution is shown by a solid line with triangles. The detailed comparison of these distributions is presented in the next section.

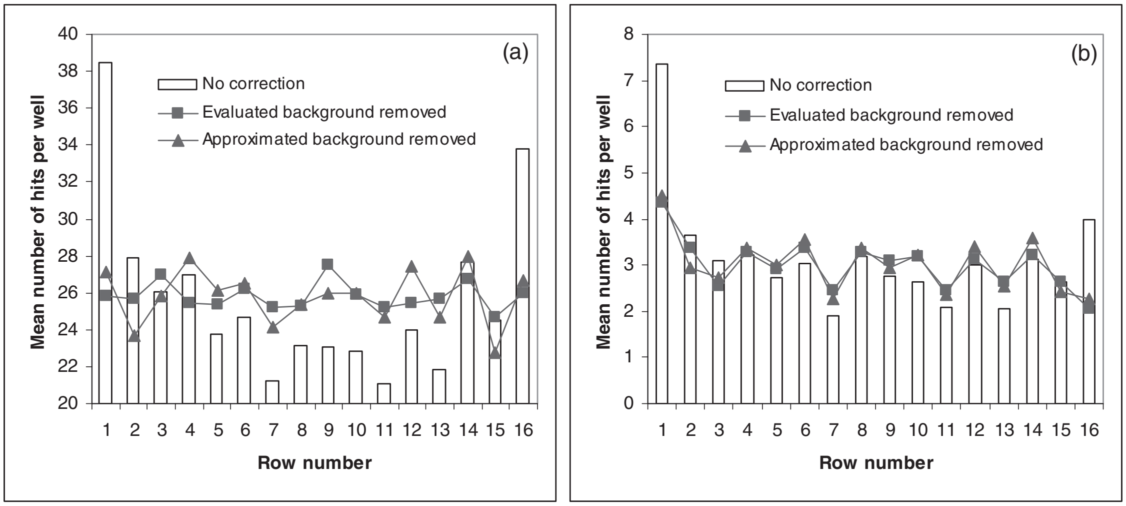

FIG. 5. Hit distribution by rows in assay 1 (164 plates): (a) hits selected with the threshold 1σ; (b) hits selected with the threshold 2σ.

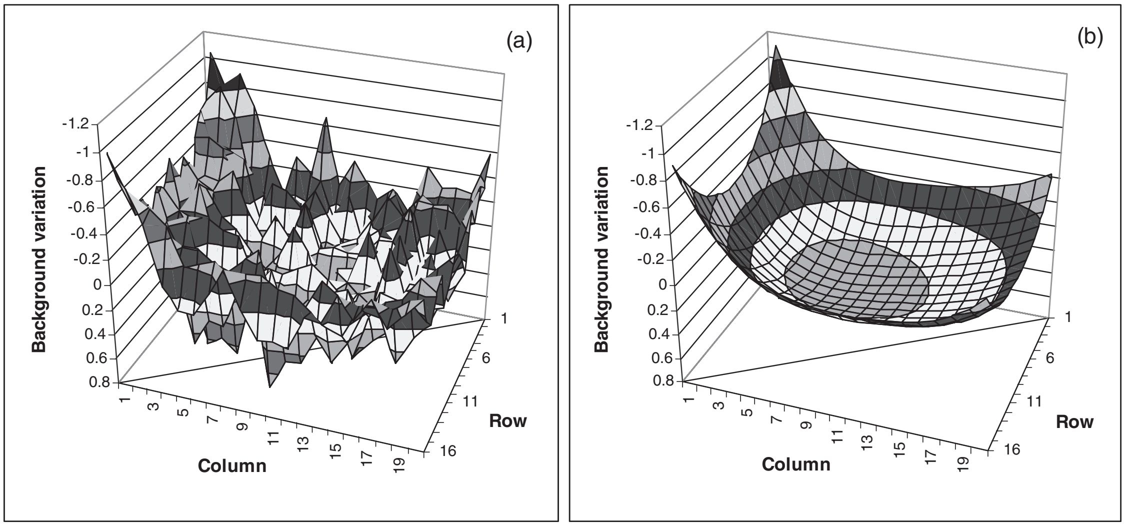

FIG. 6. (a) Evaluated background surface of assay 2 (54 plates) with 3σ hit/outlier elimination and (b) its approximation by the 4th-degree polynomial.

Assay 2

To estimate the background for the 2nd assay, comprising the data for compounds that inhibit the activity of human angiogenin, we used normalized measurements for 54 plates containing 16 rows and 20 columns each. The total number of plates in this assay was 59, but 5 of them were excluded from our analysis because of missing rows or columns. Since the number of plates was small, compared to assay 1, and due to a relatively big amount of outliers in this data set, we applied formula 6 in this case. To compute the evaluated background, we eliminated hits and outliers that deviated for more than 3σ from the mean value of each plate. Plate values without outliers were normalized with the mean centering and unit variance standardization procedure and used for the background evaluation. In Figures 6a and 6b, we present the evaluated background surface with 3σ hit/outlier elimination and its approximation by the 4th-degree polynomial. To detect the presence of systematic errors, we analyzed the hit distribution surface using the previously described procedure. We chose the hits with the 1σ selection method. The obtained hit distribution surface is presented in Figure 7a.

Most of the wells of assay 2 contain fewer than 3 hits (for 54 plates), but the hit numbers in 3 of the 4 plate corners are much

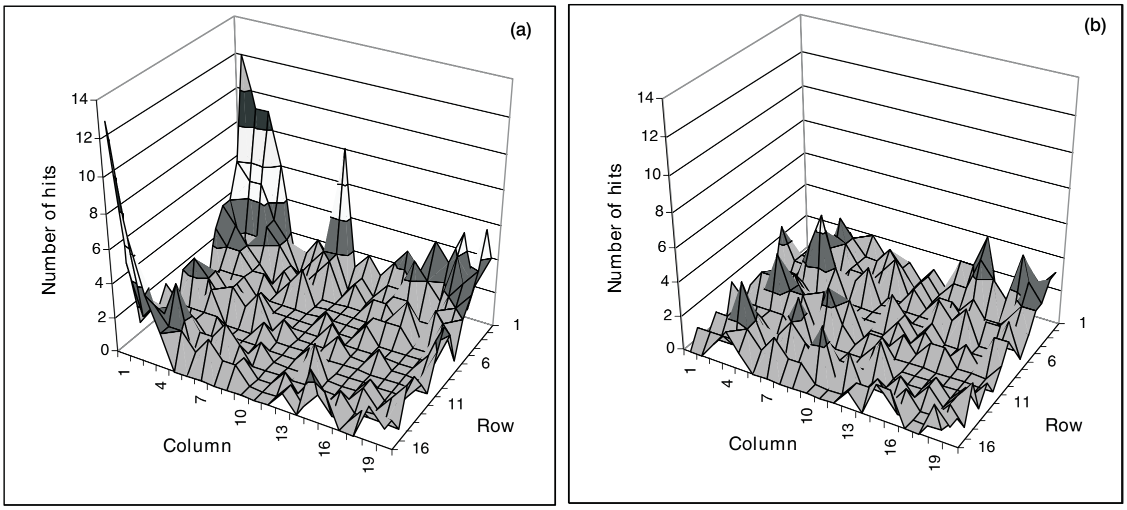

FIG. 7. Hit distribution surfaces for (a) raw and (b) corrected data sets of assay 2 (54 plates).

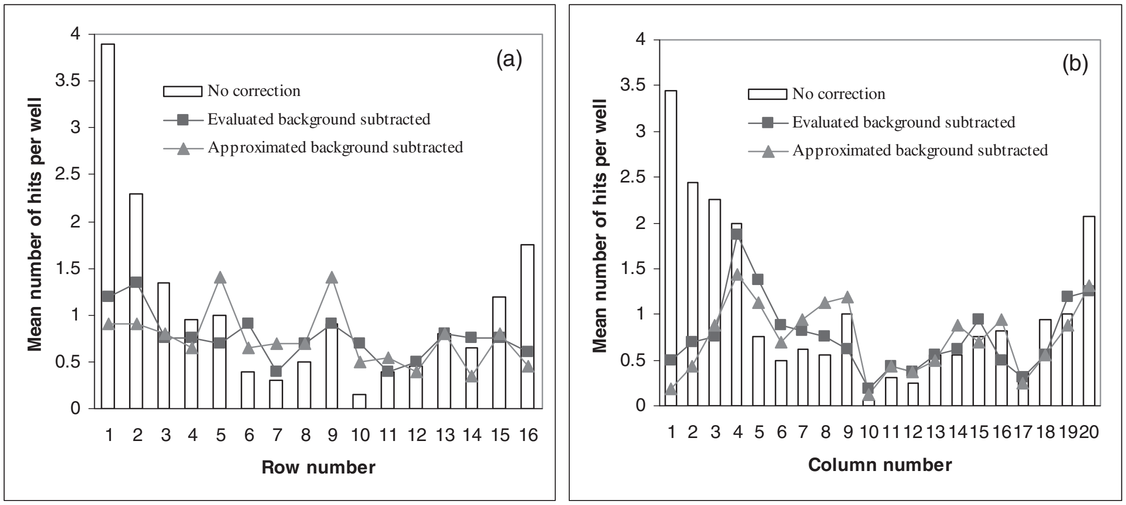

FIG. 8. Hit distribution by rows (a) and (b) columns in assay 2 for 54 plates before and after the correction.

higher. For instance, the well in column 1 and row 16 contains 13 hits. Such a big difference indicates the presence of a systematic error. The hit distribution surface illustrated in Figure 7a shows a good correlation with the evaluated background surface presented in Figure 6a.

The raw assay data were corrected to remove the impact of the systematic errors. We subtracted the evaluated background from each of the 54 plates of the assay. The corrected values were used to select the new hits. The hit distribution for the corrected data, shown in Figure 7b, is certainly more appropriate than the hit distribution of the raw data shown in Figure 7a. The high values in the plate corners (see Fig. 7a) were reduced to the reasonable values (see Fig. 7b). Also, more hits were selected (see Fig. 7b) in the low values region between the columns 4 and 19 and rows 5 and 14 (see Fig. 7a).

In Figure 8, we compared the distribution of hits by rows and columns for the raw and corrected data. The notations used in Figure 5 were also adopted here. We have to mention that the hit distribution by rows was corrected sufficiently well (see Fig. 8a), but their distribution by columns (see Fig. 8b) was still affected by local fluctuations.

We have computed the least squares deviation from the mean value for the hit distribution by rows/columns using the following formula:

$$Q = \sum_{i=1}^{n} (x_i - \mu)^2, \tag{9}$$

Table 1. Least Squares Deviations From the Mean Value for the Hit Distributions Presented in Figures 5 and 8

	Assay 1 (1σ), Fig. 5a	*Assay 1 (2σ), Fig. 5b*	*Assay 2 (1σ)*	
			Fig. 8a	*Fig. 8b*
Raw data	325.75	19.34	13.42	15.06
Evaluated background subtracted	12.73	2.97	0.94	3.13
Approximated background subtracted	34.14	4.35	1.40	2.78

where Q is the least squares deviation from the mean value, x_i is the number of hits per well at row/column i, and μ is the mean value of hits per well for the whole assay.

The least squares deviations from the mean value for the hit distributions shown in Figures 5 and 8 are presented in Table 1.

DISCUSSION

This article describes a useful method suited to analyze raw HTS data, detect systematic errors and positional effects, and make corresponding corrections of HTS signals. The method is based on the statistical analysis of signal variation within plates of an assay. It requires a sufficient amount of homogeneous experimental data to be available to produce viable results.

The normalization procedure plays an important role in this method. The application of the mean centering and unit variance standardization technique allows one to set to zero the mean values of the measurements in all plates. Regular deviations of the evaluated background from a zero plane can be explained by the presence of systematic errors. The background is computed as the mean of the normalized plate values. In the perfect case, when systematic error is absent, it leads to a zero plane. In the case of real data, the background surface will be affected by random and systematic errors. Random errors cause residuals on the background surface. They vary from plate to plate and should compensate each other during the overall background computation. Thus, their influence on the background surface can be minimized by increasing the number of plates analyzed. Systematic errors generate repeatable deviations from a zero plane. We can detect and characterize them using the trend-surface analysis of the background. The minimization of random errors certainly enhances the accuracy of the systematic error analysis. Therefore, for the small number of plates, we propose to remove outliers from the background computation.

Another problem may appear during the analysis of large assays. Plate patterns may change from batch to batch or shift over a day. We have to mention that the determination of breaks between batches should be done (preferably automatically) for large assays. In the present article, we assume that the behavior of the systematic error does not change substantially within 1 assay (see the Experimental Procedure section, assumption 5). For large industrial assays, a procedure allowing one to break (in an automated fashion) the given assay into homogeneous parts can be carried out. We can suggest 2 possible breaking procedures. First, the user could define the size of the sliding window that would go through the entire data set separating it into subsets that will be analyzed separately. Such a procedure would work under the assumption that the homogeneity is preserved for a certain interval of time. Another alternative would be to divide a large assay in k subsets (where k can be defined by the user), compute a distance measure between plates, and carry out a k-means algorithm to form k homogeneous clusters that could be analyzed separately.[11] Ideally, and we are currently working on it, the program will find an optimal k from a range of values defined by the user.

To estimate the impact of systematic errors on the hits selection procedure, we examined the distribution of hits in 2 experimental assays from the ChemBank database. The 1st one comprises experimental data for 164 plates. The screening positives are the compounds that inhibit the binding of GlcNAc to MurG. The 2nd assay includes experimental data for 54 plates. The screening positives here are compounds that inhibit the activity of human angiogenin. As hits, we marked the values that were lower than the plate means and deviated from them for more than 1σ. Brideau et al.[3] mentioned that a common method considers the threshold of 3σ for the selection of hits. Obviously, the hit selection with 1σ gave us more hits for the analysis. The common hit rate in an HTS screening campaign is in the range of 0.1% to 5%.

We have analyzed the hit distributions obtained for various σ levels. If we consider assay 2, selecting hits with the threshold of 3σ, the total number of selected hits for the whole assay will be 4. Thus, the hit rate here will be only 0.02%. The threshold of 2σ for assay 2 gives the following numbers:

- total number of selected hits = 24 and
- hit rate = 0.13%.

The threshold of 1σ for assay 2 gives the following numbers:

- total number of selected hits = 301 and
- hit rate = 1.74%.

With these numbers, the most appropriate threshold for the analysis of the hit distribution of this assay would be 1σ; it is impossible to carry out any kind of statistical analysis considering 4 or 24 hits only.

For assay 1, the threshold of 3σ gives the following numbers:

- total number of selected hits = 313 and
- hit rate = 0.54%.

The threshold of 2σ for assay 1 gives the following numbers:

- total number of selected hits = 1055 and
- hit rate = 1.82%.

The threshold of 1σ for assay 1 gives the following numbers:

Table 2. Statistical Analysis of Hit Distribution for Assays 1 and 2

Type of Correction	*Mean*	*Standard Deviation*	*Distribution by*
Assay 1 (1σ)			
Raw values (no correction)	25.69	3.752	Columns
	25.69	4.660	Rows
Evaluated background subtracted	25.83	0.953	Columns
	25.83	0.745	Rows
Approximated background subtracted	26.11	1.414	Columns
	26.11	1.522	Rows
Assay 1 (2σ)			
Raw values (no correction)	3.16	0.68	Columns
	3.16	1.25	Rows
Evaluated background subtracted	3.00	0.56	Columns
	3.00	0.54	Rows
Approximated background subtracted	3.03	0.60	Columns
	3.03	0.60	Rows
Assay 2 (1σ)			
Raw values (no correction)	1.063	0.890	Columns
	1.063	0.946	Rows
Evaluated background subtracted	0.759	0.406	Columns
	0.759	0.251	Rows
Approximated background subtracted	0.747	0.382	Columns
	0.747	0.305	Rows

- total number of selected hits = 9092 and
- hit rate = 15.7%.

Our aim was to demonstrate the impact of systematic error on the hit selection procedure. The thresholds of 1σ and 2σ gave us the best chance to show it despite a lot of false positives that were selected in these cases. It also gave us the possibility to illustrate that their impact was reduced after the data correction (see Figs. 5 and 8).

In general, the thresholds of 2.5 or 3σ would certainly be more appropriate for the hit selection, but in this case, they do not produce enough hits for our hit distribution analysis.

As shown in Figures 3, 5, and 8, a simple analysis of the hit distribution by rows and columns can indicate the presence of systematic errors. The attempt to plot the hit distribution at wells produced the complex 3-dimensional surfaces shown in Figures 4a and 7a. To identify the trends in these surfaces, we performed a trend-surface analysis discussed in this article. The comparison of the hit distribution surfaces (Fig. 4 for assay 1 and Fig. 7a for assay 2) against the background surfaces (Fig. 1 for assay 1 and Fig. 6 for assay 2) demonstrates that the distribution of hits corresponds to the background fluctuations.

To eliminate the systematic error from the original data, we subtracted the evaluated background from the raw values and reanalyzed the distribution of hits. The hit distribution by rows and columns is depicted by the solid lines with squares in Figures 5 (assay 1) and 8 (assay 2). The solid lines with triangles in Figures 5 (assay 1) and 8 (assay 2) correspond to the data corrected by the subtraction of the approximated background. The corrected data provide a more appropriate distribution by rows and by columns than the raw data (see Table 1). Both the raw and the corrected data have a comparable mean number of hits per row/column, but the standard deviation of the corrected data is 2 to 4 times lower than the standard deviation of the raw data (see Table 2). The data sets corrected by the approximated background have higher standard deviations than do those corrected by the evaluated background. This is certainly due to the small fluctuations that were not represented in the approximated surfaces.

The advantage of the proposed correction is that it is independent from a hit selection procedure. In our study, we used a 1σ deviation for the selection of hits, but the common 3σ deviation can be used as well as any other method. However, the procedures that employ high and low controls for the selection of hits will require a specific correction for the controls values. The standard approach considering controls for the hit selection is based on the following formula:

$$z_i = \frac{H - x_i}{H - L} * 100\%, \tag{10}$$

where x_i is the measured value at well *i, H* is the mean of high controls, L is the mean of low controls, and z_i is the evaluated percentage at well *i*.

The measured values of controls should be normalized along with all other values of the same plate. However, the control values must not be taken into account while computing a background surface; the controls should be considered as outliers. Subsequently, the normalized values of controls could be used in the hit selection procedure based on the following formula:

$$z'_i = \frac{H' - (x_i - b_i)}{H' - L'} * 100\%, \tag{11}$$

where x_i is the normalized value at well *i*, b_i is the background value at well *i, H′* is the mean corrected value of high controls after the subtraction of the background, L' is the mean corrected value of low controls after the subtraction of the background, and z'_i is the evaluated percentage at well *i*.

CONCLUSION

We have designed a background evaluation procedure that can be used to objectify the hit selection and provide an effective tool for the analysis and correction of HTS screens. This correction is necessary to estimate systematic errors and remove their effects from the data at hand. The described method allows one to analyze experimental HTS data and determine trends and local fluctuations of the background surface. Because the mean deviations of the background surface from a plane, computed for a sufficiently large number of plates, are caused by systematic errors, their impact can be minimized by the subtraction of the systematic background from the raw HTS data. An application of the trend-surface analysis enables one to visualize the behavior of the systematic error patterns.

In this article, we examined 2 assays of experimental HTS data from the ChemBank database. The background analyses showed the presence of systematic errors on the plate edges. We demonstrated that systematic errors can have a significant influence on the hit selection and the positional distribution of hits within plates. We corrected the HTS data for assays with 164 and 54 plates by subtracting the evaluated background from the raw data. The analysis of the corrected data sets showed that the applied modifications significantly improved the hit distribution. The positional effects caused by systematic errors were also minimized after this correction.

The software allowing researchers to carry out the background evaluation analysis of HTS data has been developed. The program is distributed as a Windows console application and its C++ source code. A graphical version of this software is freely available on our Web site (http://www.labunix.uqam.ca/makarenv/hts.html).

ACKNOWLEDGMENTS

We thank Dr. Pierre Legendre for his help in the preparation of this article. We also thank Genome Québec for funding this project.

REFERENCES

1. Heuer C, Haenel T, Prause B: A novel approach for quality control and correction of HTS data based on artificial intelligence. *The Pharmaceutical Discovery & Development Report 2003/03*. 2002. PharmaVentures Ltd. [Online]. Retrieved from http://www.worldpharmaweb.com/pdd/new/overview5.pdf
2. Gunter B, Brideau C, Pikounis B, Pajni N, Liaw A: Statistical and graphical methods for quality control determination of high-throughput screening data. *J Biomol Screen* 2003;8:624-633.
3. Brideau C, Gunter B, Pikounis W, Pajni N, Liaw A: Improved statistical methods for hit selection in high-throughput screening. *J Biomol Screen* 2003;8:634-647.
4. Heyse S: Comprehensive analysis of high-throughput screening data. *Proc SPIE* 2002;4626:535-547.
5. Zhang JH, Chung TDY, Oldenburg KR: A simple statistic parameter for use in evaluation and validation of high-throughput screening assays. *J Biomol Screen* 1999;4:67-73.
6. Zhang JH, Chung TDY, Oldenburg KR: Confirmation of primary active substances from high-throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J Comb Chem* 2000;2:258-265.
7. Helm JS, Hu Y, Chen L, Gross B, Walker S: Identification of active-site inhibitors of MurG using a generalizable, high-throughput glycosyltransferase screen. *J Am Chem Soc* 2003;125:11168-11169.
8. Kao RYT, Jenkins JL, Olson KA, Key ME, Fett JW, Shapiro R: A small-molecule inhibitor of the ribonucleolytic activity of human angiogenin that possesses antitumor activity. *Proc Natl Acad Sci U S A* 2002;99:10066-10071.
9. Lam N: Spatial interpolation methods: a review. *American Cartographer* 1983;10:129-149.
10. Legendre P, Legendre L: Trend-surface analysis. In Legendre P, Legendre L (eds): *Numerical Ecology*. 2nd English edition. Amsterdam: Elsevier Science BV, 1998:739-746.
11. MacQueen J: Some methods for classification and analysis of multivariate observations. In Le Cam LM, Neyman J (eds): *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics*. Berkeley: University of California Press, 1967.

Address reprint requests to:
Vladimir Makarenkov
Laboratoire LACIM
Université du Québec à Montréal
C.P. 8888, succursale Centre-Ville
Montréal (Québec), Canada H3C 3P8

E-mail: makarenkov.vladimir@uqam.ca