J Biomol Screen OnlineFirst, published on November 7, 2006 as doi:10.1177/1087057106293590

Using Clustering Techniques to Improve Hit Selection in High-Throughput Screening

ANDREI GAGARIN,[1] VLADIMIR MAKARENKOV,[2] and PABLO ZENTILLI[2]

A typical modern high-throughput screening (HTS) operation consists of testing thousands of chemical compounds to select active ones for future detailed examination. The authors describe 3 clustering techniques that can be used to improve the selection of active compounds (i.e., hits). They are designed to identify quality hits in the observed HTS measurements. The considered clustering techniques were first tested on simulated data and then applied to analyze the assay inhibiting *Escherichia coli* dihydrofolate reductase produced at the HTS laboratory of McMaster University. (*Journal of Biomolecular Screening* 2006:1-12)

Key words: high-throughput screening, hit selection, nonhierarchical clustering, k-means partitioning, inside-cluster distance, intercluster distance

INTRODUCTION

HIGH-THROUGHPUT SCREENING (HTS) is one of the initial stages of the drug discovery process. It allows for testing of hundreds of thousands of chemical compounds per day to select the most prominent candidates for future examination. The compounds are tested against therapeutic targets. A typical HTS procedure generates enormous data volume that requires appropriate processing tools and mechanisms. Recent developments in modern mass screening are highly influenced by the increasing number of targets identified by genomics and by the expansion of the libraries of compounds synthesized using methods of combinatorial chemistry.

Correct selection of active compounds (i.e., hits), is crucial: It is important to know that the expected drug candidate is present in the data and that the selected set of compounds does not lead to unnecessary clinical research. The mass screening process has several drawbacks including the absence of standardized data validation and the lack of reliable quality control. This makes the correct identification of active compounds quite difficult. The incoherencies are partially due to the presence of random and systematic errors in the data. Some statistical methods and software for correcting HTS data have been recently proposed in the literature. The reader is referred to the articles by Heuer et al.,[1] Gunter et al.,[2] Brideau et al.,[3] Heyse,[4] Zhang et al.,[5,6] Kevorkov and Makarenkov,[7] and Makarenkov et al.[8,9]

Another factor that influences the identification of active compounds is the hit selection procedure itself: Once the data are preprocessed and checked for quality, one has to decide which compounds should be tested in a secondary screen. However, from the statistical point of view, it is not well defined or reasonably grounded how to select the active compounds. Current practices usually apply informal rules that are based on particular laboratory constraints such as capacity limitations or financial costs of the follow-up procedures.[10]

One can identify hits by plotting raw or preprocessed measured values against compound label. First, plot the compound identity on the x axis and its activity measurement on the y axis for each plate separately and then identify compounds whose measured activity deviates from the majority of the measurements. This approach is well suited for identifying compounds with a high activity level. However, compounds of low or intermediate activity levels may be missed by such an "eyeball" procedure.

One can also select hits by computing a fixed percentage of the compounds with the highest measured activity (e.g., select 1% or 2% of the most active compounds on each plate or in the bulk of the data). This method is not statistically justified and can lead to some undesirable artifacts: The real number of hits per plate may vary a lot, and it is usually not reliable to compare measurements on different plates. Because the true number of active compounds is not known in advance, one cannot justify the selection of a fixed percentage of the primary screen compounds.

[1]Laboratoire LaCIM, [2]Département d'Informatique, Université du Québec à Montréal, C.P. 8888, succursale Centre-Ville, Montréal (Québec), Canada, H3C 3P8.

Received Apr 21, 2006, and in revised form Jul 3, 2006. Accepted for publication Jul 11, 2006.

Journal of Biomolecular Screening 11(X); 2006
DOI:10.1177/1087057106293590

Another current practice to select hits in a particular plate consists of calculating the plate mean value μ and its standard deviation σ to identify samples differing from the mean μ by at least $c\sigma$, where c is a preliminary chosen constant. For example, in the case of an inhibition assay and $c = 3$, we would select samples with measured values smaller than $\mu - 3\sigma$. This is a classical hit selection approach applied on a plate-by-plate basis. The main drawback of this hit selection procedure is the assumption that all hits have measured values smaller than a preestablished threshold depending only on the plate mean and standard deviation. To avoid this limitation, we propose using the strategy based on the following assumption: The measured values of active samples are significantly different from those of inactive ones. Such an assumption leads to a new hit selection approach that consists of finding statistically justified clusters of samples having some of the smallest or biggest measured values on each plate or in the single batch of the assay data; their values should be well distinguished from the values of all other samples. Because only a small percentage of compounds are active, the size of their cluster should be reasonably small.

Two clustering procedures will be discussed in the article. The 1st procedure considers each plate as an independent experiment, whereas the 2nd one treats all assay compounds as a single batch. The performances of the hit selection methods based on the cluster analysis with respect to the classical hit selection procedure will be illustrated first on the random data having standard normal and long-tails distributions. We will also show the differences between the 2 approaches while examining the assay inhibiting *Escherichia coli* dihydrofolate reductase generated at the HTS laboratory of McMaster University.[11]

MATERIALS AND METHODS

Random data generation and type I error

A typical HTS procedure consists of running samples arranged in 2-dimensional plates of the same format through automated screening machines that make experimental measurements. Samples are placed in wells. Most of the samples are inactive, and the measurements corresponding to active samples are assumed to be substantially different from those of inactive compounds.

We use random data sets following 2 different distributions to prove the efficiency of the cluster-based hit selection approach. The experiments were carried out on random data having the standard normal and long-tails distributions. Each random data set consists of a 1250-plate assay, with each plate having wells arranged in 8 rows and 10 columns, thus imitating the parameters of the McMaster *E. coli* screen.[11]

First, 2 random data sets with no hits were generated according to the standard normal (~N(0, 1)) and long-tails distributions. The classical hit selection procedure and 3 hit selection methods based on the cluster analysis were applied to these data. The 3 clustering strategies used to search for hits were the following: k-means partitioning, sum of the average squared inside-cluster distances (SASD), and average intercluster distance (AICD). The reader is referred to Arabie et al.[12] for an overview of clustering methods. The 3 clustering strategies considered in this study are described in more detail later in this section. Because the initial random data are not supposed to contain hits at all, the detected hits should be considered false positives. **Table 1** reports the number of false-positive hits found in the random data with no hits. Here, each plate was considered an independent experiment. In the case of classical hit selection, the sigma thresholds of $\mu - 3\sigma$ (standard normal data) and $\mu - 3.37\sigma$ (long-tails data) were applied. The sigma threshold for the long-tails data was chosen to have approximately the same number of hits that were found in the standard normal data (~120 hits in the no-hits data sets). Note that the cluster-based hit selection generally caused a small increase in the number of false positives (3.2% on average) compared to the classical hit selection.

Table 1. Type I Error: False-Positive Hits Found Using the Classical Hit Selection and the 3 Considered Clustering Methods in the Raw Random Data (With No Hits) Having Standard Normal and Long-Tails Distributions

Statistic\Distribution	*Standard Normal*	*Long-Tails*
Sigma threshold for hit selection	$\mu - 3\sigma$	$\mu - 3.37\sigma$
Number of false positives		
Classical selection	119	120
K-means partitioning	125	129
Sum of the average squared inside-cluster distances	122	125
Average intercluster distance	119	120
Interval for hit generation	$[\mu - 3.4\sigma;$ $\mu - 4.4\sigma]$	$[\mu - 4.07\sigma;$ $\mu - 5.07\sigma]$

Computations were carried out on a plate-by-plate basis.

Then, we added 1% to 5% of hits into 5 replicates of the random no-hits data. The hit locations were chosen randomly: The probability of each well to contain a hit was, respectively, 1%, 2%, 3%, 4%, and 5%. The hit values were randomly selected to be in the interval $[\mu - 3.4\sigma; \mu - 4.4\sigma]$ for the standard normal data and in the interval $[\mu - 4.07\sigma; \mu - 5.07\sigma]$ for the long-tails data, where μ denotes the mean value and σ denotes the standard deviation of the observed plate. Having data with randomly generated hits, we applied the classical hit selection procedure and the 3 considered clustering methods to detect hits in each simulated data set. The analyses were conducted for the clustering procedure working on the plate-by-plate basis and that treating all the assay data as a single batch.

Clustering methods and hit selection

One of the advantages of the clustering techniques is the possibility to find hits having values bigger than a fixed threshold (in the case of an inhibition assay). Such hits are completely ignored by the classical hit selection procedure.

In this study, we consider 3 nonhierarchical clustering methods. First, we assume that the number of clusters k is known. The objective is to partition n-given elements into the required k nonempty clusters. Clustering techniques allow objects to change their group membership through the cluster formation process. A clustering method usually starts from an initial partition chosen according to a certain criterion. Then, the reallocation of elements takes place according to an optimality criterion. Here, we consider the 3 following optimality criteria:

1. K-means partitioning[13] that minimizes the total inside-cluster variance:

$$K\text{-}means(X_1, \ldots, X_k) = \sum_{i=1}^{k} \frac{1}{N_i} \sum_{x_j \in X_i} d^2(x_j, \mu_i), \quad \sum_{i=1}^{k} N_i = n, \tag{1}$$

where X_i is the i's cluster containing $N_i = |X_i|$ elements ($N_i > 0$), x_j is an element of X_i, μ_i is the mean point of the cluster X_i, and $d(x_j, \mu_i)$ is the distance between the element x_j and the mean point μ_i of X_i.

2. Sum of the average squared inside-cluster distances between all pairs of elements (both elements of the pair must belong to the same cluster) taken over all clusters. More precisely, the method minimizes the following function:

$$SASD(X_1, \ldots, X_k) = \sum_{i=1}^{k} \frac{2}{N_i(N_i - 1)} \sum_{(x_j, x_m) \in X_i} d^2(x_j, x_m),$$
$$\sum_{i=1}^{k} N_i = n, \tag{2}$$

where X_i is the i's cluster containing $N_i = |X_i|$ elements ($N_i > 0$), x_j and x_m are 2 elements from the cluster X_i, and $d(x_j, x_m)$ is the distance between x_j and x_m.

3. Average intercluster distance between all pairs of elements belonging to different clusters (the members of each pair must be from different clusters). This partitioning method is adapted from the average linkage method of hierarchical clustering. The method consists in maximizing the following function:

$$AICD(X_1, X_2) = \frac{1}{N_1 N_2} \sum_{x_j \in X_1} \sum_{x_m \in X_2} d(x_j, x_m), \quad N_1 + N_2 = n, \tag{3}$$

where X_1 and X_2 are 2 clusters containing $N_1 = |X_1|$ and $N_2 = |X_2|$ elements, respectively, ($N_1 > 0$ and $N_2 > 0$), x_j is an element of X_1 and x_m is an element of X_2, and $d(x_j, x_m)$ is the distance between x_j and x_m.

A detailed description of the nonhierarchical partitioning methods can be found in the books of Arabie et al.[12] and Legendre and Legendre.[14]

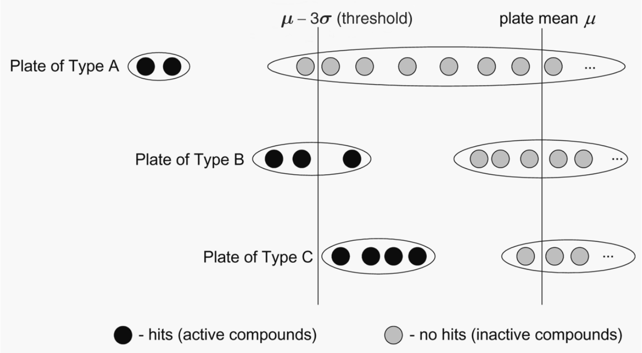
$\mu - 3\sigma$ (threshold)
plate mean μ
Plate of Type A
...
Plate of Type B
...
Plate of Type C
...
● - hits (active compounds)
○ - no hits (inactive compounds)

FIG. 1. Hit selection on the plate-by-plate basis. Three types of plates with respect to the classical hit selection threshold set to $\mu - 3\sigma$. Hits are denoted by black points, and nonactive compounds are denoted by gray points.

To be able to use a clustering method in the hit selection process, the following assumptions about HTS data in hand should be made: It is possible to divide clearly the screened samples into active and inactive, the majority of the screened samples are inactive, and measured values of the active samples differ substantially from the inactive ones. We will first show how the clustering methods can be applied on the plate-by-plate basis and then how they can be used when all the assay data are processed as a single batch.

Hit selection on the plate-by-plate basis

Following the above-mentioned assumptions, the measured values are divided into 2 groups. The bigger group contains inactive samples, and the smaller group contains active samples (i.e., hits). We assume that there is a relatively big gap that separates active samples from inactive ones. Also, we assume that it is possible to estimate the size of the gap separating the 2 groups.

We will distinguish 3 types of plates (A, B, and C) with respect to a preestablished hit selection threshold. They are depicted in **Figure 1**. All hits in the plates of type A can be found using the classical hit selection procedure, hits in the plates of type B are partially identified using the classical hit selection, and hits in the plates of type C are ignored by the classical approach. Moreover, the classical approach can also select some false positive elements. The latter elements are located outside the hit clusters in the plates of type A and close to the preestablished threshold.

An appropriate hit selection method should be able to avoid the pitfalls of the data distribution. Usually, it requires data preprocessing to ensure the accuracy of the hit selection. Ideally, active samples have similar values, form a minority group, and can be clearly distinguished from inactive ones. In the case of real data, measured values for active and inactive samples may overlap and not form 2 well-separated clusters. This can happen naturally or can be due to random and systematic errors that are usually present in HTS screens. However, despite the noise present in the measured values, it should be possible to distinguish 1 or 2 clusters of samples that have some of the smallest (or biggest) values on a particular plate.

In general, to identify clusters of hits, the following steps should be carried out for each plate independently:

- computation of the plate mean value μ and standard deviation σ (hit and outlier elimination can be also carried out at this step),
- sorting measured plate values by increasing order,
- computation of the size of the hit cluster(s) according to a chosen clustering criterion, and
- selection of hits in the obtained clusters.

Cluster construction on the plate-by-plate basis

We assume that in the absence of active compounds on a plate, there should be no significant gaps (according to a selected clustering criterion) separating the measurements. Consequently, in this case, it should be impossible to identify any kind of hit clusters in such a plate. Otherwise, hits will form a cluster (possibly divided into 2 or 3 smaller subclusters) that corresponds to some of the smallest (inhibition assay) measured values of the plate.

Let N be the number of measurements on a plate (in our simulations, N was equal to 80), and assume that the measurements are sorted by increasing order. Each of the plates can fall into 1 of the 2 possible categories:

1. some of the plate values are smaller than a fixed threshold, for example, $\mu - 3\sigma$ (plates of types A or B) or
2. all plate values are equal to or bigger than a fixed threshold (plates of type C).

Depending on the plate category, we use 2 different strategies to find the hit clusters.

Plates of types A and B. When the smallest measured value of the plate is smaller than a preestablished threshold, the 1st local minimum of the clustering function will indicate the size of a preliminary hit cluster. Because we need to partition the data into 2 groups only (i.e., active and inactive samples), the 3 clustering functions described before will be of the following form:

1. The k-means partitioning function:

$$K\text{-}means(N_1) = \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \mu_1)^2 + \frac{1}{N - N_1}\sum_{i=N_1+1}^{N}(x_i - \mu_2)^2, \quad (4)$$

where

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}x_i, \quad \text{and} \quad \mu_2 = \frac{1}{N - N_1}\sum_{i=N_1+1}^{N}x_i.$$

2. The sum of the average inside-cluster distances:

$$SASD(N_1) = \frac{2}{N_1(N_1 - 1)}\sum_{i=1}^{N_1-1}\sum_{j=i+1}^{N_1}(x_i - x_j)^2 + \frac{2}{(N - N_1)(N - N_1 - 1)}\sum_{i=N_1+1}^{N-1}\sum_{j=i+1}^{N}(x_i - x_j)^2. \quad (5)$$

3. The average intercluster distance:

$$AICD(N_1) = \frac{1}{N_1(N - N_1)}\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N}|x_i - x_j|, \quad (6)$$

where N is the total number of samples per plate, N_1 is the number of samples in the 1st cluster (i.e., hits), $N - N_1$ is the number of samples in the 2nd cluster (i.e., no hits), and x_i's are the plate-measured values sorted by increasing order.

Assume that there are at most 20% of active compounds on each plate. Then the algorithm searching for hit clusters consequently places 1, 2, 3, . . . , and $N/5$ elements ($N/5 = 16$ in our simulations) into the hit cluster and verifies whether the criterion k-means(N_1), SASD(N_1), or AICD(N_1) is increasing or decreasing. The complement to the hit cluster contains, respectively, $N - 1$, $N - 2$, $N - 3$, . . . , and $4N/5$ elements. The 1st local minimum of k-means(N_1) or SASD(N_1), or the 1st local maximum of AICD(N_1), indicates that there is a well-distinguishable gap between 2 clusters under consideration. If the 1st cluster found by 1 of the clustering methods is too big (e.g., contains more than 20% of the total number of samples of the plate), then we assume that this plate contains no hit cluster at all.

Thus, the coefficients k-means(N_1), SASD(N_1), and AICD(N_1) can be useful in finding the hit clusters containing elements whose measured values are smaller than the fixed threshold. However, if the 1st value outside the preliminary hit cluster is smaller than the fixed threshold, some hits close to the preliminary hit cluster may be ignored. In this case, we assume that the whole hit cluster is composed of 2 or more subclusters. Therefore, in this case, our program searches for the 2nd hit cluster that is indicated by the 1st local minimum of the same clustering coefficient calculated for the plate samples from which the 1st hit cluster has already been removed. The 2nd hit cluster, if found and distinguished from the remaining data, is added to the 1st hit cluster to form the final hit cluster. In some situations, the AICD function may be used to search for the 3rd hit cluster to be added to the 1st 2 to provide better selection results.

Plates of type C. Now consider plates whose values are all bigger than a preestablished threshold (**Fig. 1**). Here, the coefficients k-means(N_1), SASD(N_1), and AICD(N_1) may not be working properly: They may include too many false-positive elements into the hit clusters. Plates with a big number of hits often have all their values bigger than the preestablished threshold. This can be explained by the presence of several small values that affect the plate's mean and standard deviation. Usually, in this case, there exists a relatively big gap between the hit cluster and the no-hit values. Therefore, we assume that it is still possible to find the hit cluster by estimating this sigma-depending gap for the screen in hand.

If the sigma-gap value is underestimated, it would lead to the selection of small clusters with too many false positives and not enough true hits. An overestimation of the gap value would not allow finding the hit clusters at all. Thus, the cluster search can be very sensitive to the sigma-gap value. During the simulations, we calculated experimentally the optimal values of the sigma gaps, assuming that they depended on the average value of the maximum sigma gaps on all plates of type C. The maximum sigma-gap values between 2 consecutive elements were calculated on the 1st 20% of the plate elements. The computations were done separately for the screens with 0%, 1%, 2%, 3%, 4%, and 5% of generated hits.

Hit selection algorithm based on the cluster analysis using the plate-by-plate approach

The type of the data distribution should be determined in advance. For example, to verify the assumptions of normally distributed data, one can carry out the Kolmogorov-Smirnov test. Given a known statistical distribution, one can use the following clustering algorithm to identify hits:

- Normalize data using the zero-mean centering and unit variance standardization (also known as z-score method).
- Sort the plate measurements by increasing order.
- Compute the maximum sigma-gap value between 2 consecutive elements of the smallest 20% of the plate values for the plates of type C (**Fig. 1**), compute the average of the maximum sigma gaps for all plates of type C, and estimate the optimal sigma-gap values for the data sets with different hit percentages (see **Tables 2** and **3**).
- Compute the size of the hit cluster: For the plates of types A and B, find the 1st local minimum of the clustering coefficient k-means(N_1), SASD(N_1), or AICD(N_1) (if necessary, repeat the computation twice or thrice, removing previously found clusters); for the plates of type C, search for the cluster defined by the optimal sigma-gap value. If the cluster includes elements outside the smallest 20% of the plate measurements, disregard it (it is not a hit cluster). Also, for the plates of type C, if the cluster size is too small (e.g., less than 4% of the plate elements), disregard the cluster (it is not a hit cluster).
- Identify elements in the clusters as hits.

The optimal values of the sigma gaps calculated experimentally in the simulations based on a dichotomy search are presented in **Tables 2** and **3**. Note that in the case of no-hit data (0% of generated hits), the optimal sigma gaps are the smallest values (1.12 for the standard normal data and 1.49 for the long-tails data) that do not lead to the identification of any hit on the plates of type C. The optimal sigma-gap values are assumed to depend on the average of the maximum sigma-gap values of all plates of type C. The best-fit cubic polynomials approximating the values in **Tables 2** and **3** (see formulas 7 and 8) were calculated using the least-squares method from Maple X.[15] The main trend that may be observed in **Figure 2** is as follows: the larger the average of the maximum sigma gaps calculated on the smallest 20% of the plates' measurements, the smaller is the value of the corresponding optimal sigma-gap constant that will be used for the cluster identification.

Table 2. Correspondence between the Average of the Maximum Sigma Gaps Measured on the Smallest 20% of the Plates' Elements and the Optimal Sigma-Gap Values Used to Find Hit Clusters on the Plates of Type C for the Standard Normal Data (Calculated Experimentally)

	Generated Hits (%)					
	0	*1*	*2*	*3*	*4*	*5*
Average max sigma gap	0.447	0.453	0.53	0.71	0.82	0.85
Optimal sigma gap to be used to find clusters	1.12	0.93	0.62	0.49	0.43	0.41

Table 3. Correspondence between the Average of the Maximum Sigma Gaps Measured on the Smallest 20% of the Plates' Elements and the Optimal Sigma-Gap Values Used to Find Hit Clusters on the Plates of Type C for the Long-Tails Data (Calculated Experimentally)

	Generated Hits (%)					
	0	*1*	*2*	*3*	*4*	*5*
Average max sigma gap	0.632	0.645	0.78	0.98	1.055	1.063
Optimal sigma gap to be used to find clusters	1.49	1.04	0.76	0.56	0.5	0.48

The cubic polynomial approximating the data in **Table 2** (standard normal distribution) is as follows:

$$P(x) = 14.37 - 60.29x + 86.68x^2 - 41.28x^3. \tag{7}$$

The graph showing the correspondence of this approximation to the experimentally calculated values is represented in **Figure 2a**.

The cubic polynomial approximating the data in **Table 3** (long-tails distribution) is as follows:

$$P(x) = 22.55 - 71.54x + 78.06x^2 - 28.5x^3. \tag{8}$$

The graph showing the correspondence of this approximation to the experimentally calculated values is represented in **Figure 2b**.

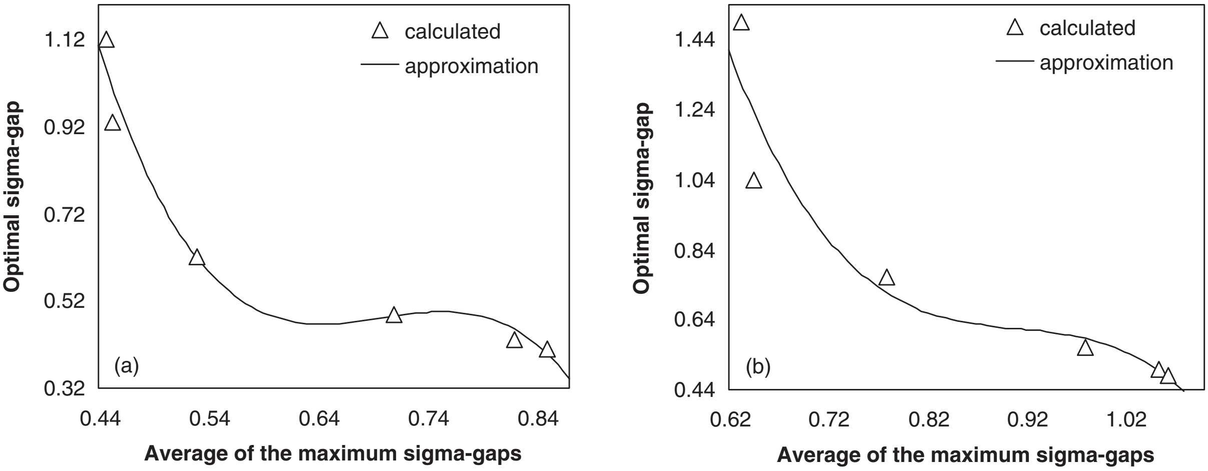

FIG. 2. Hit selection on the plate-by-plate basis. Approximation of the experimentally calculated optimal sigma-gap values by the best-fit cubic polynomial for (a) the standard normal and (b) the long-tails data.

Hit selection from a single batch

In the previous section, we considered the hit selection process that treats each plate as an independent experiment. Here, we focus on the analysis considering the whole assay data as a single batch. To conduct this kind of experiment, we first have to make sure that all plates of our assay were processed under the same testing conditions. This analysis can be recommended when a well-optimized assay protocol shows little plate-to-plate variability.

Thus, we conducted the simulations treating altogether the data of the whole assay. Two algorithmic strategies are possible in this case. First, we can still process the data on the plate-by-plate basis but use the parameters (here, the mean value and standard deviation) computed for the data from the whole assay instead of those computed for each particular plate. Such a strategy will not be sensitive at all to the data distribution (i.e., the data can be distributed randomly or not). Thus, the compounds that are not randomly distributed can be processed in this way. No important changes in the above-presented algorithm should be done to implement this strategy.

The 2nd strategy, the strategy that we actually tested, assumes that all the assay data are coming from a large single plate. Thus, the 3 considered clustering procedures can be tested in turn to find the best, according to the selected criterion, partition of the entire data set into the clusters of active compounds (i.e., hits) and inactive compounds. We assume that this separation should occur not far from the traditional hit selection threshold $\mu - 3\sigma$, where μ is the mean value of the whole assay and σ is the assay standard deviation. This strategy proceeds by testing a fixed number of cluster partitions and selects the partition that optimizes the value of the given clustering coefficient (see **Fig. 3**). Up to 500 cluster partitions (i.e., the search interval in **Fig. 3**) were tested in our simulation study for each considered random data set.

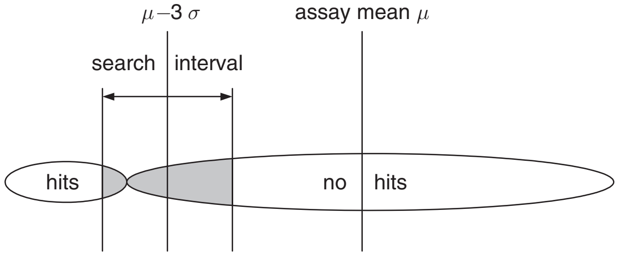

FIG. 3. Hit selection from a single batch. Classical hit selection procedure selects as hits the compounds whose values are lower than $\mu - 3\sigma$. The clustering procedures look for the best partitioning of the given assay into the clusters of hits and no hits. The gray area indicates the search interval for partitioning.

RESULTS AND DISCUSSION

Experimental results for the hit selection carried out on the plate-by-plate basis

In this study, we examined 2 random HTS assays. Both simulated assays consisted of 1250 plates having wells arranged in 8 rows and 10 columns. The measurements of the 1st assay followed a standard normal distribution (~N(0, 1)), and the measurements of the 2nd assay followed a long-tails distribution. First, we conducted the analysis on the plate-by-plate basis. Numerical results for the classical hit selection and the 3 clustering methods considered in this article are presented in **Tables 4** and **5**. The statistics reported in **Tables 4** and **5** were obtained by running the simulation program 100

Table 4. Hit Selection on the Plate-by-Plate Basis: Average Hit Selection Results for the Standard Normal Data Obtained Using the Classical Hit Selection Procedure and the 3 Clustering Methods

	Generated Hits (%)				
	1	*2*	*3*	*4*	*5*
Generated hits (number)	998.5	1997.1	2997.6	3992.5	4994.6
Classical hit selection					
Sigma threshold	3.0σ	2.88σ	2.82σ	2.76σ	2.67σ
Total hits found	1069.1	1930.4	2521.3	2872.4	3180.7
False positives	106.5	125.6	126.3	115.4	107.3
False negatives	37.1	192.2	602.5	1235.5	1919.6
Correct hit rate (%)	96.3	90.4	79.9	69.1	61.6
K-means partitioning					
Total hits found	1119.2	2096.1	3108.2	4079.2	5041.2
False positives	127.9	136.4	135.7	125.8	116.4
False negatives	12.3	26.2	27.6	34.8	57.9
Correct hit rate (%)	98.8	98.7	99.1	99.1	98.8
Sum of the average squared inside-cluster distances					
Total hits found	1098.4	2083.8	3061.9	4046.1	4978.8
False positives	117.9	123.3	123.7	113.9	106.9
False negatives	13.8	38.4	54	72.9	104.8
Correct hit rate (%)	98.6	98.1	98.2	98.2	97.9
Average intercluster distance					
Total hits found	1077.8	1987.4	2845.9	3708.3	4629
False positives	106.9	102.5	96.2	84.2	78.7
False negatives	27.2	118.7	247.3	366.7	446
Correct hit rate (%)	97.3	94.1	91.8	90.8	91.1

Table 5. Hit Selection on the Plate-by-Plate Basis: Average Hit Selection Results for the Long-Tails Data Obtained Using the Classical Hit Selection Procedure and the 3 Clustering Methods

	Generated Hits (%)				
	1	*2*	*3*	*4*	*5*
Generated hits (number)	999.8	2005.3	2997.9	3996.5	4994.7
Classical hit selection					
Sigma threshold	3.37σ	3.27σ	3.19σ	3.16σ	3.09σ
Total hits found	1047.9	1764.8	2134.8	2132.1	2090.3
False positives	103.5	123.2	118.3	89.2	73.5
False negatives	56.3	361.8	984.7	1953.2	2980.2
Correct hit rate (%)	94.4	82	67.2	51.2	40.4
K-means partitioning					
Total hits found	1105.1	2099.7	3108.7	4079.1	5056.1
False positives	125.2	129.5	118.7	93.9	77.1
False negatives	18.2	31.8	7.5	9.2	15.5
Correct hit rate (%)	98.2	98.4	99.8	99.8	99.7
Sum of the average squared inside-cluster distances					
Total hits found	1099.3	2080.4	3082.2	4061.3	5016.1
False positives	117.9	122.3	114.5	86.8	73.1
False negatives	20.5	35.6	16	20.4	28.8
Correct hit rate (%)	98	98.2	99.5	99.5	99.4
Average intercluster distance					
Total hits found	1084.6	2041.5	2999.1	3931.7	4903.9
False positives	112.8	113.8	105.2	78.4	64.6
False negatives	27.7	82	106.6	145.4	161.4
Correct hit rate (%)	97.2	95.9	96.5	96.4	96.8

times for each random data set and calculating the average values of the runs.

In the case of 1% hit data, the hits were classically selected by choosing values lower than the thresholds $\mu - 3\sigma$ and $\mu - 3.37\sigma$ for the standard normal and long-tails data, respectively. In the case of other hit percentages, the classical sigma threshold was adjusted to keep the number of false positives close to that found by the clustering methods. Note that a lower threshold increases the number of false negatives. This is due to a trade-off between the number of true hits and the number of false negative hits (see **Fig. 4a**).

Thus, we computed the true hit detection rate, the false-positive (i.e., inactive samples that were identified as hits) rate, and the false-negative (i.e., real hits that were not detected) rate during the simulations. The true hit detection rate was much higher for all 3 clustering methods compared to the classical hit selection procedure (see **Fig. 5**). For both data distributions, k-means partitioning and SASD clustering outperformed AICD clustering and the classical hit selection procedure.

The classical hit selection implies a trade-off between the true hit rate and the false-positive rate. The influence of a fixed threshold on the false-positive and false-negative rates is illustrated in **Figure 4**. This figure shows a schematic distribution and an overlap between hit and no-hit measurements with respect to a fixed classical threshold (**Fig. 4a**). **Figure 4b** shows the distribution and the overlap between the false-positive and false-negative values in the threshold area for the generated standard normal random data with 5% of added hits. An increase in the number of correctly found hits obtained by adjusting the classical threshold usually implies a sensitive increase in the number of false positives. The cluster approach can attenuate this artifact: Hit clusters are not sensitive to any preestablished threshold and can grab more correct hits without an important increase in the false positive and false negative rates. For both standard normal and long-tails data including 1% to 3% of generated hits (see **Tables 4** and **5**), the number of false positives remained at the same level as the number of false positives in the raw data with no hits (in our case, approximately 120 false positives).

Note that the k-means partitioning method detected slightly more true hits than the SASD method did (see **Fig. 5**). However, the SASD method performed slightly better than k-means partitioning in terms of false positives (see **Tables 4** and **5**). The AICD method was certainly the best in terms of

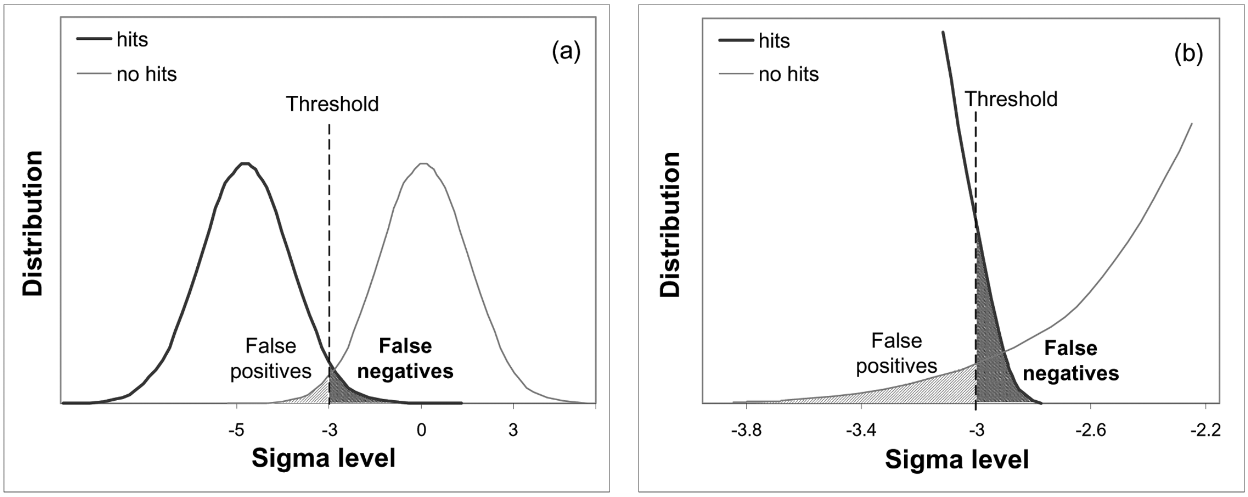

FIG. 4. Influence of a fixed threshold on the false positive (gray area) and false negative (black area) rates in the case of standard normal data. (a) Typical distribution of hits and no hits. (b) Distribution of hits and no hits in the threshold area for the generated standard normal random data with 5% of added hits.

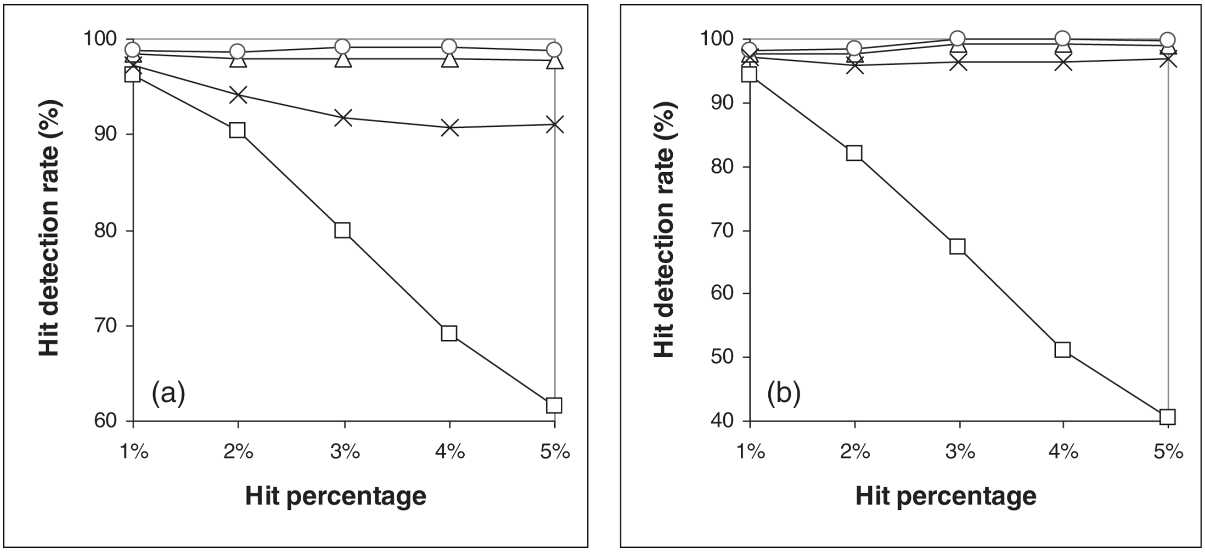

FIG. 5. Hit selection on the plate-by-plate basis. Variation of the true hit detection rate depending on the hit percentage. (a) Standard normal data. (b) Long-tails data. Hit selection methods: □ = classical; ○ = k-means partitioning; △ = sum of the average squared inside-cluster distances clustering; × = average intercluster distance clustering.

false positives but very inconsistent in terms of false negatives (see **Tables 4** and **5**).

Experimental results for the hit selection carried out for a single batch of data

We also carried out simulations to test the procedure processing all the assay data at the same time. Similarly to the plate-by-plate approach, we conducted our experiments on the two 1250-plate assays having standard normal (~N(0, 1)) and long-tails distributions of data. The obtained results for the 4 competing strategies, including the traditional hit selection method, are given in **Tables 6** and **7**. The reported statistics were obtained as average results obtained after 100 runs. Similar to the simulation described in the previous paragraph, the hit percentage varied from 1% to 5%.

The hits were randomly generated with respect to the above-described procedure. Note that the sigma thresholds reported in **Tables 6** and **7** were different from those reported in **Tables 4** and **5**. Here, they were adjusted in a way that the classical hit

Table 6. Hit Selection from a Single Batch: Average Hit Selection Results for the Standard Normal Data Obtained Using the Classical Hit Selection Procedure and the 3 Clustering Methods

	Generated Hits (%)				
	1	*2*	*3*	*4*	*5*
Generated hits (number)	1004.5	2002.9	2994.4	3992.2	5004.9
Classical hit selection					
Sigma threshold	3.0σ	2.58σ	2.30σ	2.10σ	1.94σ
Total hits found	1043.0	1998.1	2957.0	3941.5	5007.7
False positives	60.9	203.9	399.7	643.0	965.0
False negatives	22.5	208.7	437.1	693.8	962.2
Correct hit rate (%)	97.8	89.6	85.4	82.6	80.8
K-means partitioning					
Total hits found	1038.0	2097.1	3156.0	4140.5	5306.7
False positives	60.3	234.9	447.9	710.7	1078.6
False negatives	26.8	140.8	286.3	562.5	776.8
Correct hit rate (%)	97.3	92.3	90.4	85.9	84.5
Sum of the average squared inside-cluster distances					
Total hits found	1091.1	2097.1	3156.0	4140.5	5003.7
False positives	86.5	234.9	447.9	710.7	963.7
False negatives	0.0	140.8	286.3	562.5	964.9
Correct hit rate (%)	100.0	93.0	90.4	85.9	80.7
Average intercluster distance					
Total hits found	1039.0	1994.1	2953.0	3937.5	5003.7
False positives	64.5	202.5	398.9	641.7	963.7
False negatives	26.0	211.3	440.3	696.4	964.9
Correct hit rate (%)	97.0	89.5	85.3	82.6	80.7

Table 7. Hit Selection from a Single Batch: Average Hit Selection Results for the Long-Tails Data Obtained Using the Classical Hit Selection Procedure and the 3 Clustering Methods

	Generated Hits (%)				
	1	*2*	*3*	*4*	*5*
Generated hits (number)	1001.6	2004.2	3001.4	3991.6	4999.1
Classical hit selection					
Sigma threshold	3.12σ	2.70σ	2.38σ	2.10σ	1.95σ
Total hits found	985.6	1940.6	2896.5	4069.5	4912.9
False positives	203.1	540.6	883.7	1378.6	1636.7
False negatives	219.1	604.2	988.5	1300.7	1722.9
Correct hit rate (%)	78.1	69.9	67.1	67.4	65.5
K-means partitioning					
Total hits found	1084.6	2039.6	3095.5	4368.5	5511.9
False positives	251.1	570.6	960.8	1470.8	1845.9
False negatives	168.0	535.1	866.6	1093.9	1333.1
Correct hit rate (%)	83.2	73.3	71.1	72.6	73.3
Sum of the average squared inside-cluster distances					
Total hits found	1084.6	2039.6	3095.5	4368.5	4908.9
False positives	251.1	570.6	960.8	1470.8	1635.2
False negatives	168.0	535.1	866.6	1093.9	1725.4
Correct hit rate (%)	83.2	73.3	71.1	72.6	65.5
Average intercluster distance					
Total hits found	976.6	1936.6	2892.5	4065.5	4908.9
False positives	198.9	539.4	882.2	1377.3	1635.2
False negatives	223.9	607.0	991.0	1303.4	1725.4
Correct hit rate (%)	77.6	69.7	67.0	67.3	65.5

selection procedure selects the number of hits close to the real number of generated hits. Because of this difference, the hit detection rate for the classical hit selection procedure reported in **Tables 4** and **6** and **Tables 5** and **7**, and illustrated in **Figures 5** and **6**, respectively, cannot be actually compared between them. The 3 clustering algorithms were then carried out within the search interval (see **Fig. 3**) in the area of the classical hit selection threshold. The search intervals with 10, 100, 250, and 500 elements were tested in our study, and the best results were reported. In general (see **Fig. 6**), the k-means partitioning and SASD clustering procedures outperformed AICD clustering and the classical hit selection method.

One can notice that the results of the clustering methods obtained using the plate-by-plate approach (**Tables 4** and **5**) are better, in almost all cases, than those obtained using the single-batch approach (**Tables 6** and **7**). This can be explained by the fact that the latter approach is dependable on the classical hit selection threshold and does not offer the possibility of studying independently the plate distributions of measured values as the plate-by-plate approach does.

Searching for hits in the experimental data

We applied the classical hit selection procedure and the 3 considered clustering methods to the experimental data set generated at the McMaster University HTS laboratory and compared the obtained results. This HTS assay is publicly available at the following Web site: http://hts.mcmaster.ca/HTSDataMiningCompetition.htm (see also the work of Zolli-Juran et al[11]). It consists of a screen of compounds inhibiting *E. coli* dihydrofolate reductase. Each compound was screened twice: 2 copies of 625 plates were run through the screening machines. This gives 1250 plates in total, each having wells arranged in 8 rows and 12 columns (columns 1 and 12 containing controls were not considered in this study). The assay conditions reported in Zolli-Juran et al[11] were the following: Assays were carried out at 25° C and performed in duplicate. Each 200-µL reaction mixture contained 40 µM NADPH, 30 µM DHF, 5 nM DHFR, 50 mM Tris (pH 7.5), 0.01% (w/v) Triton, and 10 mM β-mercaptoethanol. Test compounds from the screening library were added to the reaction before initiation by enzyme and at a final concentration

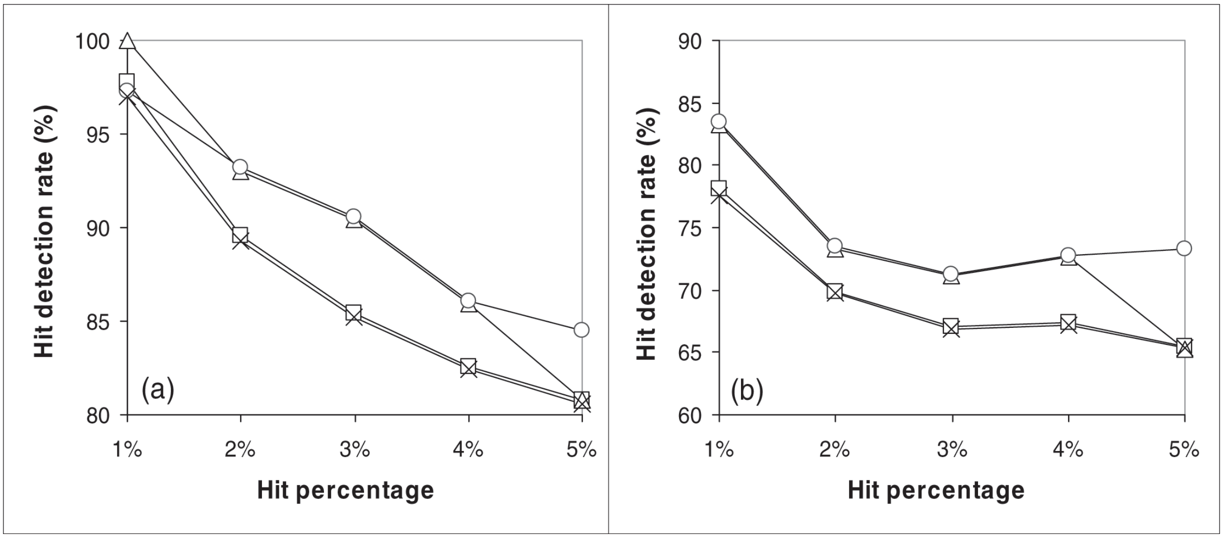

FIG. 6. Hit selection from a single batch. Variation of the true hit detection rate depending on the hit percentage. (a) Standard normal data. (b) Long-tails data. Hit selection methods: □ = classical; ○ = k-means partitioning; △ = sum of the average squared inside-cluster distances clustering; × = average intercluster distance clustering.

of 10 µM. All data are reported as the percentage residual activity relative to the average of the high controls.

The fact that the original study has identified only 32 hits in both copies (for more detail see, http://hts.mcmaster.ca/Competition_FAQ.html) shows that either some kind of random noise was added to the data during the analysis or that the testing conditions were slightly different for the 2 assay replicates. This could also happen because the plate-to-plate variability was high or the testing conditions were inconsistent from one replicate to another. Thus, we decided to carry out the plate-to-plate clustering analysis that is more appropriate than the hit selection from 2 batches, 1 per replicate, in such a situation.

The distribution function for this data set and its approximation by a Gaussian distribution are shown in **Figure 7**. It is worth noting that for this representation, the experimental data were plate normalized using the zero-mean centering and unit variance standardization. The Gaussian distribution was modeled using the parameters of the experimental data. The classical hit selection threshold was set to $\mu - 3\sigma$. The upper right corner of **Figure 7** shows the data distribution in the hit selection area.

To apply properly a clustering method, it is first necessary to calculate the average of the maximum sigma gaps for the plates of type C (**Fig. 1**). There were precisely 886 plates of type C in the McMaster data set. The average of the maximum sigma gaps on the smallest 20% of elements of these plates was equal to 0.485. Using the approximation by the polynomial (formula 7), we obtained the value of the optimal sigma-gap constant to be used for identifying hit clusters on the plates of type C. The sigma-gap constant was equal to 0.81 in this case.

The classical hit selection procedure found 429 hits in the McMaster data, the k-means partitioning method found 467 hits, the SASD method found 465 hits, and the AICD method found 757 hits. The bar chart representing the hit selection results is shown in **Figure 8a**. The k-means partitioning method detected 39 hits that were not identified as hits by the classical hit selection procedure (see **Fig. 8b**), the SASD method found 38 extra hits, and the AICD method found 328 extra hits. On the other hand, almost all hits detected by the classical procedure were confirmed by the clustering methods: The k-means method missed only 1 classical hit, the SASD method missed only 2 hits, and the AICD method did not miss any of the classical hits. The intersections between the sets of hits provided by the 3 considered clustering strategies are

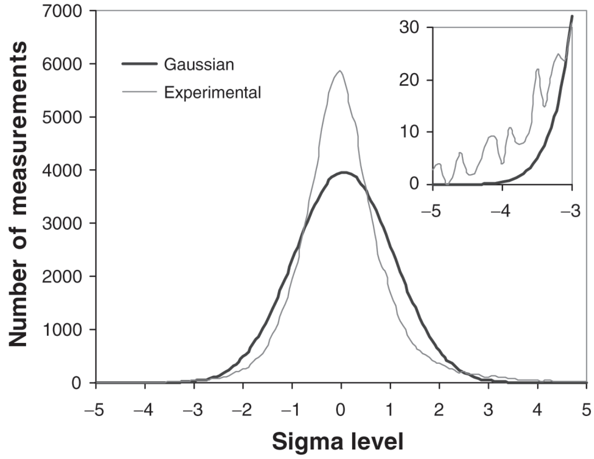

FIG. 7. Distribution of measurements at the McMaster *Escherichia coli* assay (1250 plates) and its comparison to a Gaussian distribution.

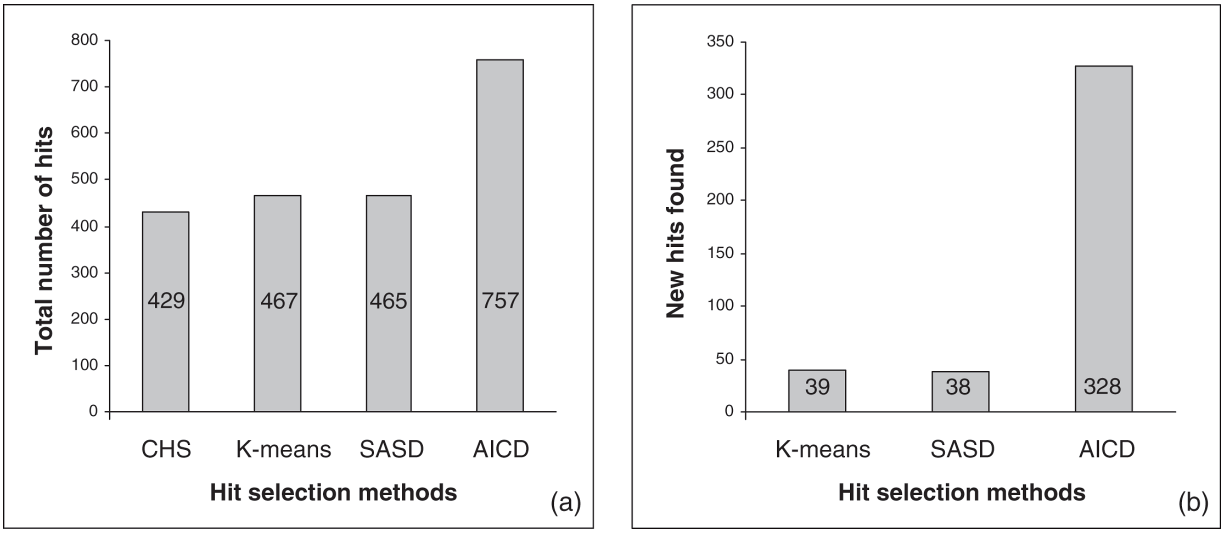

FIG. 8. Comparison of the results provided by 4 hit selection methods for the considered McMaster University experimental HTS screen. (a) Total number of selected hits. (b) Number of hits found by the 3 clustering methods and not found by the classical hit selection procedure.

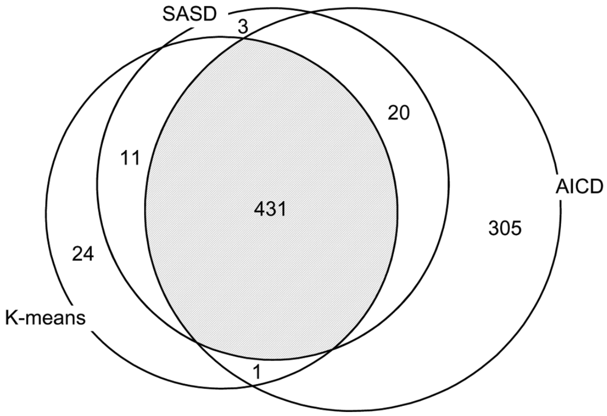

FIG. 9. Intersections between the 3 hit sets found by the 3 considered clustering methods for the McMaster University experimental screen.

shown in **Figure 9**. Note that the results obtained by k-means partitioning and SASD were very similar, whereas the AICD procedure found 305 hits that were not detected by any other clustering method.

Let us compare in more detail the results provided by the classical hit selection procedure and those obtained by the SASD method. The classical procedure identified 429 hits in the McMaster assay, whereas the SASD method found 465 hits in the same data set (total increase of 36 hits). Note that 26 of these 36 hits were found on the plates of type C (**Fig. 1**); that is, their values as well as all values in their clusters were bigger than the classical threshold of $\mu - 3\sigma$. These 26 hits were found in the clusters containing 3 and 4 elements located on 8 plates. Only 2 hits detected by the classical approach were not identified as hits by SASD. These 2 elements were located on the plates of type A (**Fig. 1**) that had nonempty hit clusters. Both values of the nondetected classical hits were close to the threshold of $\mu - 3\sigma$. Thus, it is likely that these 2 elements identified as hits by the classical method were false positives. The SASD method also found 12 hit elements with measured values bigger than $\mu - 3\sigma$ located in the clusters on the plates of type B (**Fig. 1**).

Note that each compound of the considered McMaster University assay had 2 copies and thus was tested twice during this screen (for more detail, see the screen description on the McMaster University Web site). Both the classical hit selection procedure and SASD selected 36 compounds were confirmed as hits for both copies of the compound. However, the 36 compounds confirmed twice by the classical procedure were different in 1 compound from the 36 compounds confirmed twice via SASD.

CONCLUSION

We described a new approach to select hits in HTS data. The presented approach is based on the cluster analysis of assay measurements. We considered 3 clustering techniques that enabled us to improve classical hit selection results in the simulations with random data. Two clustering schemes were examined, with the 1st one considering each plate as an independent experiment and the 2nd one processing all the data as a single batch. We agree with Malo et al[10] that improving hit specificity and sensitivity cannot be met by technological and organizational improvements alone and that improvements in data analysis methods are needed to fulfill the promise of HTS.

The results of the considered clustering techniques depend on the data distribution as well as on the plate size and the number of plates. Given an experimental HTS data set, we recommend trying clustering methods on the random data that have the identical distribution and are arranged in the same number of plates of the same size. The random data should be generated and modeled using the mean values and standard deviations of the experimental data. This will allow one to choose plausible clustering methods and parameters for hit selection in the experimental data. The simulations with random data can be done by analogy with the computational experiments described in this article.

Based on the simulations with random data, we recommend using k-means partitioning or the SASD clustering method. In general, the 2 methods have shown better performances than the AICD method and classical hit selection. However, if it is more important to have a low number of false positives in a particular HTS assay, the AICD method can be considered as well. It is worth noting that it is possible to combine the clustering hit selection methods with data correction methods and the classical hit selection. Moreover, one can also combine the hit selection methods, for example, searching first for the initial hit cluster using the k-means partitioning method and then applying the AICD method to the remaining plate elements (in this study, the k-means partitioning was the best method in terms of true hits, and AICD showed the best performance in terms of false positives). A more conservative option would consist of the selection of compounds that were identified as hits by all considered clustering methods. It also would be interesting to test the popular k-medoids method[16] in the hit selection context and to incorporate the available chemical information about the compounds into the clustering methods (e.g., molecular weight, reactivity level, etc.). This would lead to multivariable data sets that provide the possibility of using weighting variables.

The experiments described in this article showed that the application of different clustering techniques leads to different hit selection results. Therefore, it would be interesting to design and carry out significance tests for the hit selection methods. One can also simulate and analyze some other types of data to confirm the advantages of the cluster-based hit selection.

Finally, we also suggest trying methods of machine learning that would combine the obtained information on experimental HTS data with chemical description parameters of the tested compounds. Such a combination of quantitative and qualitative descriptors seems to be very promising for an efficient selection of high-quality drug candidates.

ACKNOWLEDGMENTS

We thank Genome Quebec for funding this project. We are also thankful to Dr. Robert Nadon (McGill University, Montreal, Canada) for his help with the data simulations. We are also thankful to an anonymous referee for her or his helpful comments.

REFERENCES

1. Heuer C, Haenel T, Prause B: A novel approach for quality control and correction of HTS data based on artificial intelligence. In *Pharmaceutical Discovery & Development Report 2003/03*. Oxford, UK: PharmaVentures Ltd, 2002.
2. Gunter B, Brideau C, Pikounis B, 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, 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. *Proceedings of 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. Kevorkov D, Makarenkov V: Statistical analysis of systematic errors in high-throughput screening. *J Biomol Screen* 2005;10:557-567.
8. Makarenkov V, Kevorkov D, Gagarin A, Zentilli P, Malo N, Nadon R: *An Efficient Method for the Detection and Elimination of Systematic Error in High-Throughput Screening*. Unpublished manuscript, 2006.
9. Makarenkov V, Kevorkov D, Zentilli P, Gagarin A, Malo N, Nadon R: HTS-Corrector: software for the statistical analysis and correction of experimental high-throughput screening data, *Bionformatics* 2006;22:1408-1409.
10. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R: Statistical practice in high-throughput screening data analysis. *Nat Biotechnol* 2006;24: 167-175.
11. Zolli-Juran M, Cechetto JD, Hartlen R, Daigle DM, Brown ED: High throughput screening identifies novel inhibitors of *Escherichia coli* dihydrofolate reductase that are competitive with dihydrofolate. *Bioorg Med Chem Lett* 2003;13:2493-2496.
12. Arabie P, Hubert LJ, De Soete G: *Clustering and Classification*. Hackensack<N> NJ: World Scientific, 1996.
13. 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*. Vol. 1, Statistics. Berkeley: University of California Press, 1967.
14. Legendre P, Legendre L: *Numerical Ecology*. 2nd ed. Amsterdam: Elsevier Science BV, 1998.
15. Char BW, Geddes KO, Gonnet GH, Monagan MB, Watt SM: *Maple Reference Manual.* New York: Springer-Verlag, 1988.
16. Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

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

E-mail: gagarin@lacim.uqam.ca