

### Introduction

High throughput screening experiments are typically used within the pharmaceutical industry for the identification of candidate drugs. Using a high throughput screen with automated fluorescent imaging platform allows a large number of chemical compounds to be tested in a biological assay in order to identify any activity inhibiting or activating a biological process. The images produced by the screen contain a wealth of information that can be used to define fully the effects of a compound on cells and therefore have an advantage over conventional *in vitro* screening techniques. It is for this reason the fluorescent imaging assays have been termed high content screening [1].

The study analysed here involves use of an automated robotic system to administer compounds to cellular assays and take measurements of various aspects of cell activity from a high content image (example images can be seen in Figure 1). These images are analysed and quantified using advanced imaging algorithms to produce a set of variables.

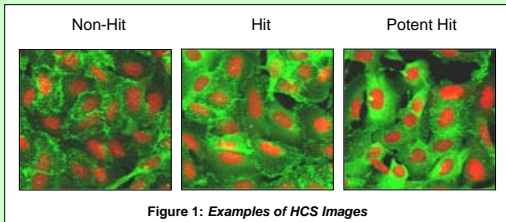


Figure 1: Examples of HCS Images

In order to sample a diverse a chemical space as possible, each high content screen may extend to a million or more individual assays [2]. This, and the fact that only a small number of compounds in a screen (<1%) exhibit the desired biological effect leads to a number of statistical challenges in data analysis.

Classification methods are used in this analysis to pinpoint compounds that may have the potential to be developed into future drugs (denoted as **hits**). False positives (i.e. compounds that on inspection prove to be misclassified as hits) are denoted as **false hits** and those that do not exhibit the desired biological effect are denoted **non-hits**. Examples images of hits, non-hits and false hits can be seen in Figures 1 & 2.

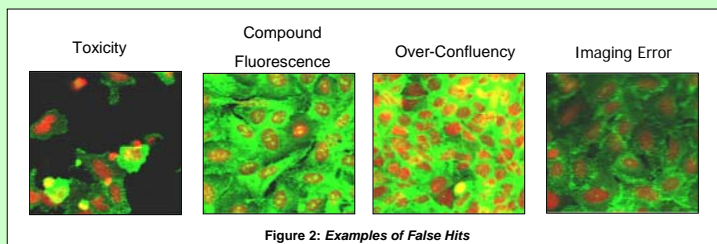
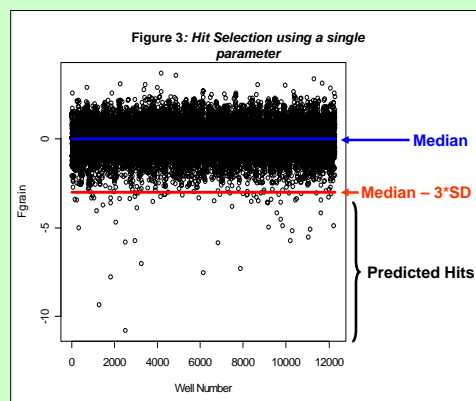


Figure 2: Examples of False Hits

### Data

The data available from the screening experiment were collected in three batches. The first of 12,285 compounds were selected because of their known properties and were used in a pre-screen to validate the experimental procedures. The data from this batch form the training data. The remaining two batches of 33,941 and 33,408 compounds (labelled A and B respectively) yielded the test data for classifying each of these 67,000 compounds. 15 variables were measured for each compound, each was derived as averages over all cells of individual image properties.



### Current Methodology

Compounds are currently classified using a single parameter selected as the most sensitive during assay development. Hits are identified as those compounds with values more than 3 standard deviations below the median.

However, a low value may also occur from false positives so all compounds initially predicted as hits are visually checked and classified as **[true] hits**, **potent hits**, or **false hits**. The one parameter hit selection procedure can be seen in Figure 3.

Approaches for multi-parametric classification are still in their infancy but it is believed that proper exploitation of the information contained in within each high content screen image will enable more refined compound selection. This leads to the following objectives:

- develop a multi-parametric approach to enable more refined compound selection;
- reduce the number of false positives to reduce the 'cost' of manual image inspection;
- avoid an increase in the number of false negatives since these are potential new drugs.

### General Statistical Objective:

To develop a multi-parametric classifier to enable a more refined approach to compound selection in high content screening experiments.

### Preliminary Analysis

The first four columns of Figure 5 show the true classifications of predicted hits for batches A and B combined using the single parameter approach, linear discriminant analysis and random forests. This preliminary analysis shows that the latter two multi-parameter approaches misclassify fewer compounds when predicting true hits. In particular, linear discriminant analysis has the lowest misclassification rate of the three methods. However, note that the random forest identifies more hits than either of the other two methods. This suggests that this classifier is worth pursuing further and refining to lower the misclassification rate.

The results of the preliminary analysis combined with the fact that the data in the training set were selected because of their known properties highlights the key problem:

### Key Problem:

Traditional multivariate classification methods assume the training data is a random sample from the same distribution as the test data

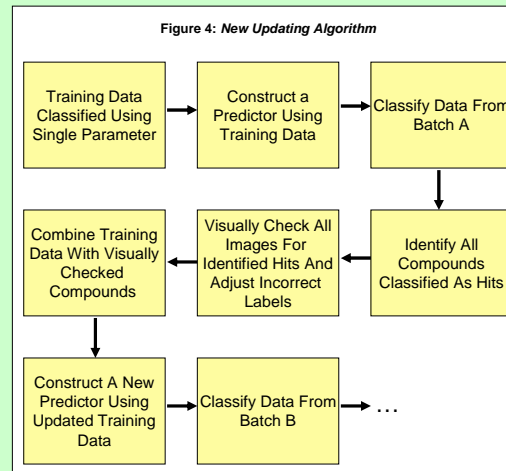
**BUT**

Training data from high content screens are selected because of their known biological effects

We propose a new method for updating classification rules as a solution to this problem.

### New Updating Algorithm

The methodology for the new updating algorithm is as follows. The training data is initially classified using the single parameter plus visual checking approach described previously and a random forest predictor is constructed using these data. This predictor is then used to classify those compounds which were screened as part of batch A into groups of true hits, non-hits and false hits.



The compounds identified as true hits by the random forest predictor are examined visually to verify the predictions. At this stage all true hits that have been misclassified are corrected. A new training data set is now created by combining the data from batch 1 with the visually checked compounds from batch A.

This new updated training data is used to construct a new random forest predictor for the classification of Batch B. This part of the algorithm accommodates the possibility that the training data is not representative of the test data by correcting the assumptions on underlying distributions made from the training data.

For each new batch of data the training data is updated using the previous batches until the final predictor that is constructed is the 'best' possible. At this stage it is recommended that each of the batches are classified again to see if any true hits were missed during previous classification. This procedure is illustrated in Figure 4.

### Application

Figure 5 shows the combined results of classifying batches A and B using the updating algorithm with a random forest as the classifier. The results shown are the true classifications of the compounds that were selected as hits. This approach is used because it is not possible to find the true classification of all 67,349 compounds. Additionally, the results of classifying using the one parameter approach, linear discriminant analysis and non-updated random forest are also shown.

Figure 5: Classification of Predicted Hits

	1 Parameter	Linear Discriminant Analysis	Random Forest	Updated Random Forest
Hit	69	73	96	<b>97</b>
Potent Hit	51	50	49	<b>49</b>
Non-Hit	0	52	50	<b>1</b>
Focus Error	31	3	7	<b>0</b>
High Background	21	2	6	<b>0</b>
Over Confluent	5	14	45	<b>0</b>
Toxic	10	2	25	<b>0</b>
Well Dry	4	0	1	<b>0</b>
No Visible Image	10	0	0	<b>0</b>
Low Draq5	3	0	0	<b>0</b>
% Misclassified	<b>41.18 %</b>	<b>37.24 %</b>	<b>48.03 %</b>	<b>0.68 %</b>

### Results

The results of classifying the two batches of data suggest that the new methodology performs better than both the single parameter approach and the two other multi-parameter approaches. The most noticeable improvement is the reduction in the number of false positives, with the updating method only misclassifying 1 compound out of the 147 selected. However, the updating method does fail to identify two of the potent hits that were selected by the single parameter approach.

### Discussion

The results of applying this algorithm to data show that the initial objectives have been met. The proposed iterative updating multi-parametric methodology substantially increases the overall number of true hits found whilst dramatically reducing the number of false positives. However, it should be noted that since non-hits in the training set are never visually checked the number of false negatives is unknown. A focus of current work is to extend the methodology to pinpoint those non-hits which are most plausibly false negatives, thus increasing the number of hits identified further.

### References:

- [1] P.A. Clemons (2004). Complex phenotypic assays in high-throughput screening. *Curr. Opin. Chem. Biol.* **8**, 334-338.
- [2] B.A. Kenny, M. Bushfield, D.J. Parry-Smith, S. Fogart, J.M. Treherne (1998). The Application of High-Throughput Screening to Novel Lead Discovery. *Prog. Drug Res.* **51**, 245-269.
- [3] N. Malo, J.A. Hanley, S. Cerquozzi, J. Pelletier, R. Nadon (2006). Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167-175.