Data to Knowledge in Pharmaceutical Research

Dr. Ann DeWitt[*],
Saziye Bayram[†], German Enciso[‡], Harshini Fernando[§],
Justin Kao[¶], Bernardo Pagnoncelli[‖], Deena Schmidt[**],
and Jaffar Ali Shahul Hameed[††]

August 18, 2004

Abstract

This report is concerned with the analysis of data from "high-throughput" screening of possible drug compounds. High-throughput screening is a relatively new process yielding thousands of data points at a time, more than can be handled by traditional methods of biological data analysis. We examine a few methods for extracting knowledge from this data and also illustrate the use of descriptors for predicting drug activity. Finally, we present suggestions for improvements in the process and ideas for future work.

1 Introduction

The lengthy process of bringing pharmaceutical products from concept to market begins with drug discovery. A significant part of modern drug discovery is the testing of thousands of compounds in a chemical library for drug-like activity. One of the primary tools of this testing is high-throughput screening, a highly automated system to assess the biological activity of thousands of compounds at a time.

The main purpose of HTS is to find chemical families that have the desired activity and to show a structure-activity relationship. A secondary purpose of HTS is to elucidate biological insights given multiple types of biological results, initiating further wet lab experimentation. Because high-throughput screening is a manufacturing process with highly variable output and active compounds comprise only a small fraction of the

[*] 3M Corporation
[†] State University of New York, Buffalo
[‡] Rutgers University
[§] Texas Tech University
[¶] Northwestern University
[‖] Pontifícia Universidade Católica, Rio de Janeiro
[**] Cornell University
[††] Mississippi State University

compounds tested, one challenge in pharmaceutical research is extracting knowledge from the large amounts of data generated.

In this report, we illustrate part of the process of converting data to information. We

1. describe the data set we were given
2. investigate two methods of analyzing the high-throughput data, comparing their results to a low-throughput reference
3. study possible descriptors for predicting drug activity
4. discuss possible extensions of approaches and additional approaches that could be the focus of future work.

2 Experimental methods

Biological screening can generally be categorized as "low-throughput" or "high-throughput". Low-throughput screening (LTS) is a bench-scale assay of compounds by chemists or biologists in a controlled laboratory environment, often involving several experiments with the same compound at different concentrations or under other carefully varied conditions. Because of the small number of compounds involved (less than 100), low-throughput assays are typically interpreted manually by scientists.

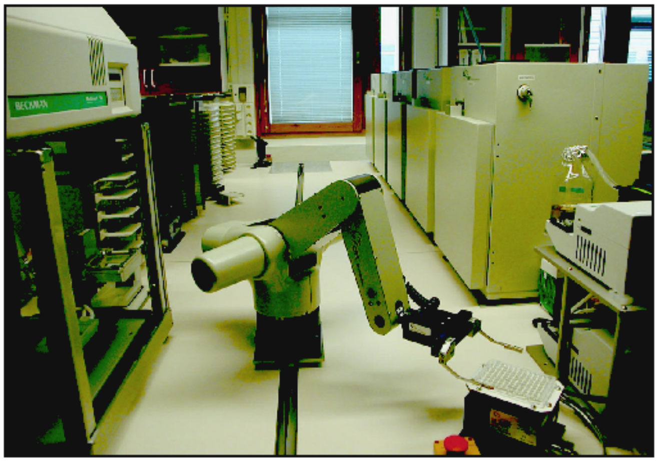

Figure 1: A high-throughput screening robot. Photo from [4].

By contrast, high-throughput screening (HTS) is an automated procedure for quickly testing many compounds at once. In the type of HTS considered here, chemists prepare "plates" consisting of many small wells, each with a compound to be tested. These are placed in the input queue of a robot which performs the actual experiements, adding cells or other reagents and recording the resultant activity signal (e.g., luminence). Due to the large volume of data generated and the manufacturing-type conditions, computer processing is usually required to understand HTS data.

In summary, HTS is used to screen compound libraries for biological activities while LTS is necessary for developing appropriate experimental protocols and providing the expert-interpreted results to which HTS is compared.

3 Data

Our first task was to analyze results from one assay (#1) conducted in both LTS and HTS to find chemical compounds that induce "B cell" proliferation. In this activation assay, active compounds cause high signals whereas inactive compounds generate low signals. The data in Assay #1 includes several controls: a natural positive reference known to induce B cell growth, a natural negative reference known to inhibit B cell growth, positive and negative references with structural similarity to tested compounds, and a (negative) solvent control, DMSO, with no added compound.

Also, we had four other assays which, combined with the first, targeted biological and structure-activity relationships. Unlike Assay #1 in which we were given essentially raw data, this group of assays was given to us in an analyzed form so that we could explore various biological predictors for active compounds.

3.1 Assay #1: low-thoughput screen

The low-throughput activity screen consists of 31 experiments testing a total of 64 compounds to give 1272 rows of data. The data set includes four controls: natural positive reference, positive reference, negative reference and the solvent DMSO control. In each experiment, one or more compounds are tested at concentrations between 0.1 μM and 30 μM. This gives us a dose-response curve for each compound in each experiment. The data set also includes expert testimony on the results as to which compounds are active or inactive.

Low-throughput is a more detailed data set for each individual compound than high-throughput since each compound is tested for activity at several concentrations. However fewer molecules are tested—64 in LTS versus 1614 in HTS (with an overlap of only ~ 40 compounds)—so scientists are typically able to assess the assay results manually due to the small number of compounds tested per experiment. Because of these factors, LTS data is perceieved as more reliable in terms of determining active molecules than HTS data. Therefore, once we have a ranking of active compounds for LTS, we make it the standard by which to validate HTS rankings.

3.2 Assay #1: high-throughput screen

The high-throughput screen uses standard industry plate formats, either 96 wells/plate or 384 wells/plate for a total of 33 plates. This gives about 6000 wells in which 1614 compounds are tested. The screen includes all controls listed at the beginning of this section. A typical plate is organized

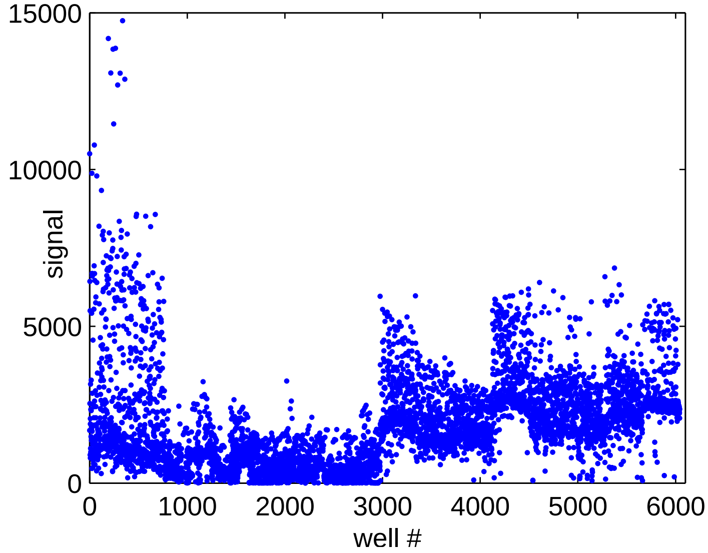

Figure 2: Raw HTS data.

so that controls appear in the first two columns in 4-replicates and test compounds cover the rest of the plate. Some plates also have controls scattered throughout, but not in any regular manner. As can be seen in Figures 2-3, there are many systematic biases and a large amount of noise in the raw data.

3.3 Assays #2-5

Once a screen has been analyzed, an important next step in the drug development process is to provide information to chemists and biologists in the form of structure-activity relationships or new hypotheses about biology. We examined four assays and discerned relationships between Assay #1 and other assays (#2,#3,#5). The compounds were organized into 10 different chemical structure groups. Whereas Assay #1 measures B cell proliferation, Assays #2 and #3 measure protein concentration, and Assays #4 and #5 measure compound-receptor activation. Assays #2 and #3 are reported in terms of "potency", the minimum protein concentration necessary for a signal to be detected above background noise, and "efficacy", the maximum protein concentration seen in a dose-response curve. A brief explanation of each screen is given as follows:

Assay #2 - Peripheral blood mononuclear cell screen (PBMCs) PBMCs in tissue culture media are placed in plate wells pre-dispensed with compound, and 24 hours later the media is inspected for the concentration of two proteins (Protein "1" and Protein "2"). Protein 2 is known to activate a particular signaling pathway through Receptor "2" and is thus thought to be indicative of B cell prolif-

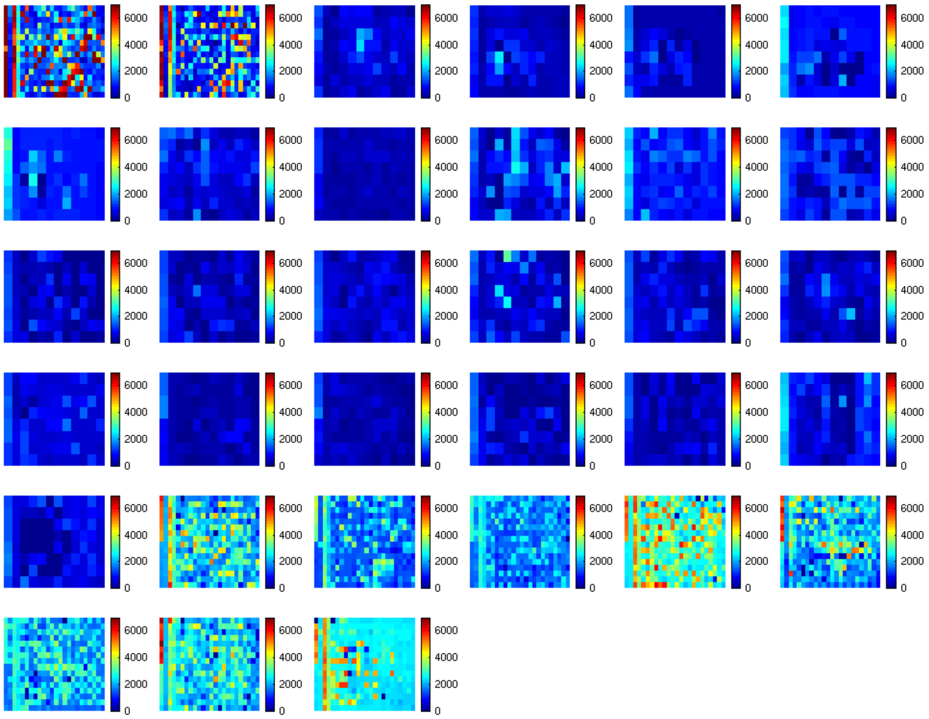

Figure 3: Raw HTS data by plates. Red indicates higher recorded signal.

eration. Although Protein 1 is seen in the assay, it is unclear what relationship exists between Protein 1 and B cell proliferation, although scientists suspect one exists. We assume Protein 2 is the most relevant one. The outcomes of this screen are potency and efficacy scores.

Assay #3 - D screen "D cells" in tissue culture media are placed in plate wells pre-dispensed with compound, and 24 hours later the media is inspected for the concentration of two proteins (Protein 2 and Protein 3). Again, Protein 2 is thought to be indicative of B cell proliferation because it activates Receptor 2's signaling pathway. Protein 3 is not expected to have a direct relationship to B cell proliferation.

Assay #4 - Activation of Receptor 1 Compounds were investigated for their ability to activate Receptor 1. The results are classified as agonist if a compound activates a receptor and indeterminate or not seen, if not. We didn't consider this assay in our analysis due to time constraints.

Assay #5 - Activation of Receptor 2 Compounds were evaluated for their ability to activate Receptor 2. The results are classified as in Assay #4. Agonist for Receptor 2 is thought to be indicative of B cell proliferation and upstream of Protein 2.

4 Analysis

Once a library of compounds has been run through HTS, it is required to analyze the data to find the active compounds, or "hits". Some of the difficulties involved include high signal-to-noise ratios, inherent variability of biological targets, minor malfunctions in the instrumentation, and systematic biases due to environmental effects.

We describe the spectrum of approaches to extracting this knowledge—heuristic data analysis, statistical modeling of the HTS data, process modeling of the HTS, and phenomenological modeling of the biology. For this project, we illustrate the first two approaches and simply discuss the others, due to lack of time and information.

4.1 LTS Reference ranking

Because the low-throughput screening is a much more reliable and comprehensive data set for the compounds that it includes, it is used as a standard for evaluating our HTS results. Here we describe the methods used to analyze the LTS data. Although this analysis was a prerequisite for the HTS analysis, it was quite different in purpose—the goal was to produce a standard for HTS analysis.

First, in order to compare results across the 31 different experiments in the LTS, it is necessary to normalize the data. We tried several different methods,

1. normalizing signals on a scale of negative reference (=0) to positive reference (=1),

2. calculating the n-fold increase of signals over the solvent (DMSO) control,
3. and calculating the n-fold increase of signals over the negative reference.

The third method is significantly more straightforward than the first, and was expected to be more accurate than the n-fold increase over the DMSO because the solvent control was only tested once per experiement. (Moreover, the negative reference is known not to be fully inactive so this provides a more stringent means to identify active compounds.) Indeed, according to expert opinion, the results produced by the last method matched their understanding almost exactly. This involved calculating (signal)/(negative reference) at all concentrations tested and taking the median value over all experiments.

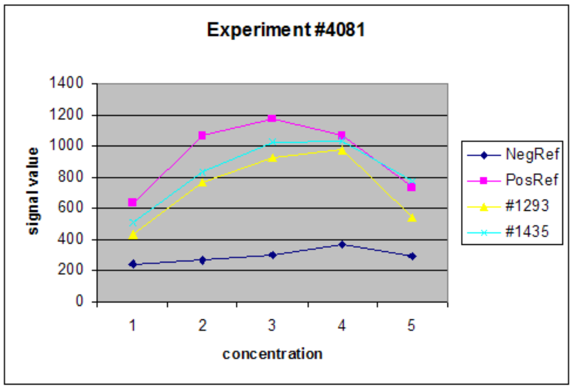

Figure 4: A typical dose-response.

Next, we used the numbers thereby obtained to rank our compounds in order of activity. For active compounds, we expect signal values to increase with increasing concentration (e.g., Figure 4 for a typical dose-response curve). This is usually the case, but we observe many signal values decreasing at 30 μM. One reason for this effect is thought to be cytotoxicity at high concentrations of some compounds. By taking the median of our numbers, we prevent such data points at the end of the curve from skewing our ranking.

We use expert testimony to reconcile ambiguities between potentially active versus inactive compounds and choose a cutoff for hits. The expert advised us to remove experiment #4035 due to unreliable data, i.e. the negative reference signal was higher than the positive reference signal at concentration 0.3 μM. See Figure 5. Based on this advice, we excluded three more experiments: experiment #3787 also showed signal values where negative reference > positive reference and two experiments (#3433 and #3460) showed negative signal values.

The end result of the LTS analysis is a ranking of compounds and “hit/non-hit” determinations which can be used as a baseline for evaluat-

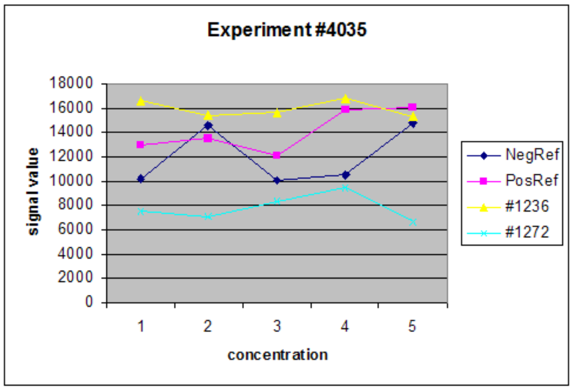

Figure 5: A flawed dose-response.

ing HTS analysis. (The match to expert opinion was best using an n-fold increase cutoff value of 1.145, giving 33 active compounds.)

4.2 Heuristic data analysis

The simplest and most common methods for finding hits in HTS data are by heuristics. These are commonly-used guidelines for quality control of data, normalizing plates, and scoring compounds. It is both an advantage and drawback of heuristics that they do not need to describe the errors being corrected for—there is minimal accounting for sources of variability in the signal. Some popular heuristics are the use of Z-scores (a statistical parameter) to normalize the data by making each plate's standard deviation and median comparable, and the use of controls to verify the reliability of the data on a plate. These are frequently used in industry to evaluate HTS data, and we discuss below the results of our heuristic analysis and its validation.

We applied a three-stage heuristic algorithm to our data:

1. Remove plates using fixed criteria derived from expert opinion.
2. Normalize signals using plate-by-plate Z-scores.
3. Rank each compound by the median of its signals and mark as hits those with signals exceeding one standard deviation of the screen median.

Plates were excluded from our data set if any of the mean, median and trimmed mean of the negative references on the plate were greater than the corresponding statistics for the postive references, or if any two of these statistics for the solvent control (DMSO) were larger than those for the data. These criteria were derived from expert opinion in removing experiments from the LTS screen. A total of 10 plates were excluded from our analysis by these criteria (leaving roughly 200 out of 1600 compounds without any data).

The Z-score was calculated by the following formula for each data point:

$$z_i = \frac{x_i - \tilde{x}}{\sigma_x},$$

where $\tilde{x}$ is the median and σ_x is the standard deviaton of the signal values on a particular plate.

The last step consists of grouping the Z-scores compound by compound, taking the median value and sorting these values. This yields a ranking of the compounds, in order of predicted drug activity. We also have an alternate "ranking" of hit/non-hit, which is determined by taking as hits those compounds for which a signal exceeds the screen median by at least one standard deviation. Using the binary hit/non-hit system, we arrive at at 72.7% match with the LTS results. In other words, of the compounds included in both screens, 72.7% of them were classified in the same way. A more qualitative evaluation is shown in Figure 6, which depicts the LTS ranking vs. HTS ranking of common compounds. If the rankings were perfectly matched, this would be a monotonically increasing curve (Figure 6a) or a diagonal line (Figure 6b).

4.3 Statistical modeling

A more systematic method for finding hits is statistical modeling — using a portion of the data to derive statistical predictors that determine hits in the rest of the data. This is a purely mathematical technique that does not rely on specific information about the HTS mechanism or the biology involved.

In particular, we used the compounds identified by LTS as training data for the following algorithm, derived from [7]:

1. Calculate a variety of binary parameters for each data point (e.g., *"Is the positive reference on its plate at least one standard deviation above the data median?"*).
2. Identify parameters that have predictive value by comparing the set of hits and the set of non-hits. Create a "kernel" from these parameters.
3. Calculate the Hamming distance between every data point and this kernel.
4. Identify points as hits if they are at least as close to the kernel as the mean distance of the training data.

The end result is a division of the HTS compounds into two sets, hits and non-hits. Unfortunately, due to the relatively small size of the LTS data set and the lack of an independently verifiable classification (e.g., synthetic data), we are unable to definitively evaluate the results here. However, due to the small number of parameters identified in the kernel (8), it is believed that the training data is not over-fitted. In particular, there is a 78.4% match with the heuristic results for HTS (out of 1607 compounds), and a 78.8% match with the LTS results (out of 33 compounds). A qualitative ranking plot (where the statistical model's compound ranking is based on Hamming distance to the kernel) is shown in Figures 7-8.

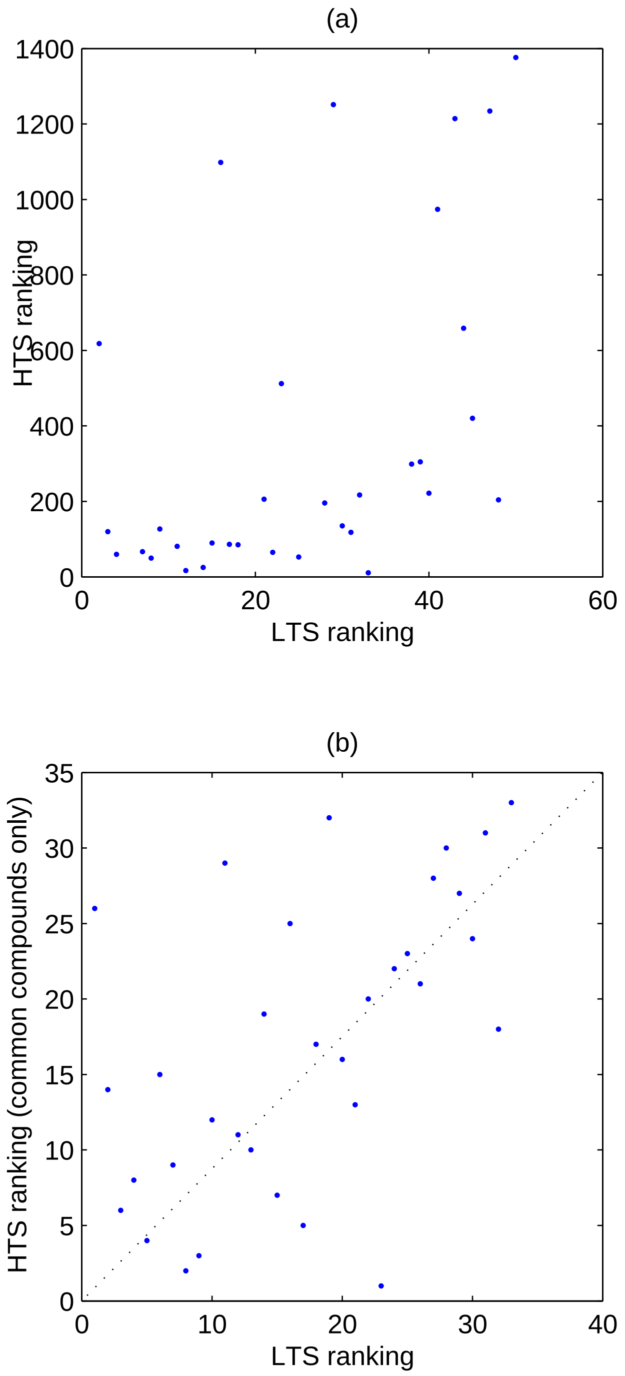

Figure 6: Heuristic HTS ranking vs. LTS ranking. (a) Compounds in both rankings are shown. Rank 0 is most active. (b) Same as (a), but HTS rankings are relative.

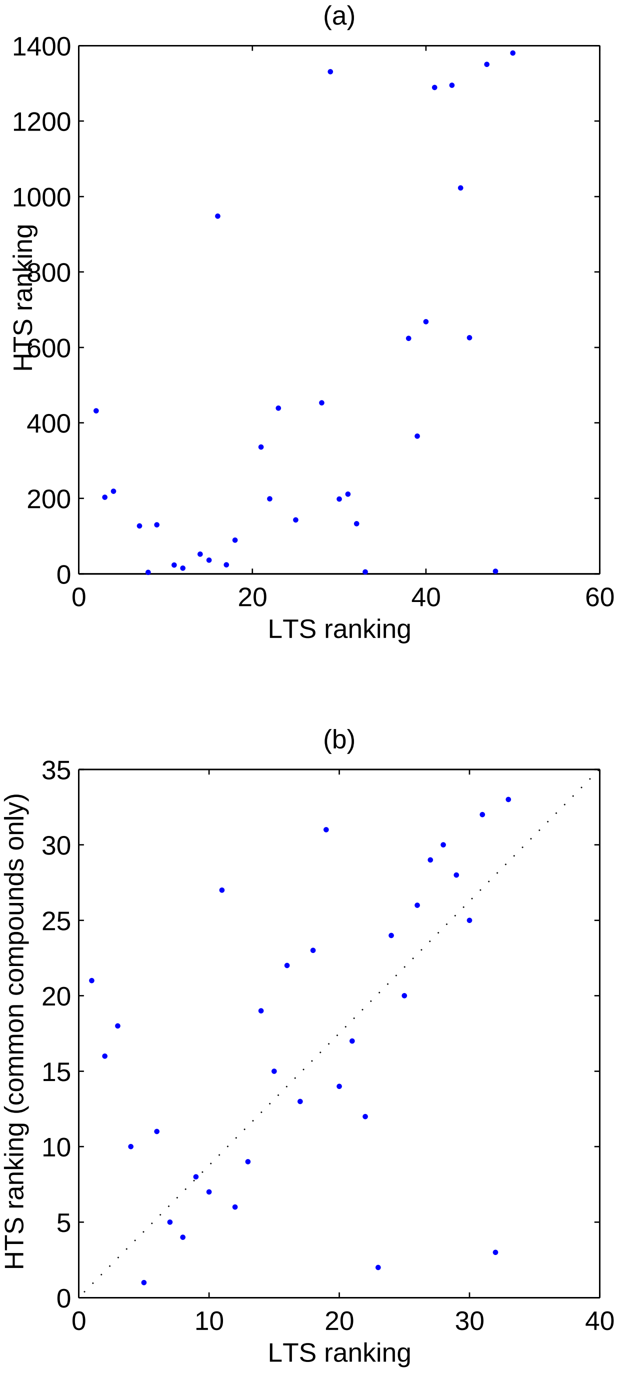

Figure 7: Statistical model HTS ranking vs. LTS ranking.

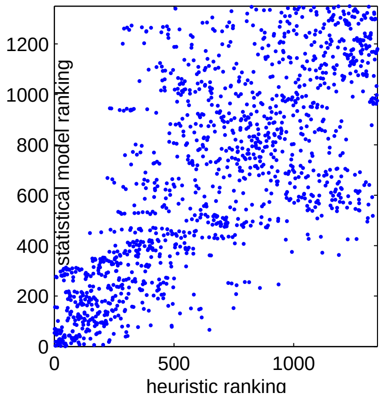

Figure 8: Statistical model HTS ranking vs. LTS ranking.

4.4 HTS modeling

Another approach is modeling of the HTS process itself, particularly the factors contributing to variation in the signal. Some of these might be: well position, actual compound concentration in a given well, cytotoxicity of the compound, environmental factors such as temperature and humidity, and plate run order. This would take the form of a functional relationship

$$y_i = f(x_i, \mathbf{z}_i) + \epsilon,$$

where y_i is the observed signal for each compound i, x_i is the actual activity, $\mathbf{z}_i$ are environmental variables and ϵ is noise in the system. A good approximation for f would allow one to invert the relationship and approximate the activity x_i, given y_i. While a model of this sort would be ideal, we unfortunately did not have sufficient information about the HTS process in this project to constuct such a model.

4.5 Biological modeling

At the most specific level, there is modeling of the actual biological processes involved. At the cellular scale, this sometimes takes the form of a reaction-diffusion-advection system of differential equations describing the concentrations of chemicals or proteins and their effect on cell processes. Another approach is modeling gene activity as a network of logical elements. However, modeling cell biology is not feasible for HTS screening, due to the wide variety of compounds tested — the complexity of the resulting system would have been prohibitive, and beyond our knowledge of biology. Inversely, HTS is not appropriate for biological modeling because the information generated is not sufficient. Had we been studying the behavior of a specific drug-target system, this would be an appropriate approach.

5 Recommendations for the HTS process

When analyzing the HTS assay we encountered insufficient and/or biased information. For example, some signal values of the compounds are unexpectedly high compared to other signal values within a plate. As another example, the negative reference signal is greater than the positive reference signal on a few plates. We suggest the following improvements to the high throughput screening (HTS) process:

Increase the number of samples per compound The major obstacle we found when trying to create a model for the HTS was the small number of samples per compound. For a significant statistical analysis of each compound we suggest running a larger number of experiments with each compound. In our data we have only 2 or 3 samples per compound, which makes model construction impractical. If we had more samples per compound, we could model the noisy signals as a normal distribution.

Robotic error analysis Behavior and possible error of the instrumentation should be characterized. Despite the precision of HTS, we should be aware of possible inconsistencies and/or measurent errors.

Improve plate consistency The use of two different sized plates is an unnecessary inconsistency. A variation in the plate size introduces one more variable to the problem without any apparent gain. In future experiments we suggest a uniform size of the plates in the HTS so that the data could be better analyzed.

Environmental effects on data We assume that signal of the compounds is affected by environmental parameters such as temperature, humidity, pressure and light. These are possible causes of the abnormal signal values we obtained, but we did not have this information for our analysis. This would have been useful in making our results more accurate, and in constructing more realistic and descriptive models.

6 Biological and structure-activity relationships

Once hits from an HTS campaign have been identified, the results can be used to find relationships between a set of chemical or biological descriptors and drug activity. We show the application of statistical models and heuristics to other sets of experimental data in order to uncover these relationships. Such results can be used by chemists to pursue more potentially active compounds, or by biologists to formulate new hypotheses and drive wet lab experimentation.

6.1 Heuristics

We carried out a preliminary analysis using heuristic algorithms. These results were then compared with the HTS screen for activity. In the analysis of Assay #2 (PBMCs), it is thought that higher values of potency and efficacy are required for a compound to be active. We used the product of potency and efficacy to identify potentially active compounds, taking the mean of these products as a reference. A similar analysis was carried out on Assay #3 (D cells).

A different kind of analysis was performed on Assays #4 and #5. We investigated which compounds activate Receptor 1 and Receptor 2. Assay #4 was omitted from the analysis in the interest of time. Furthermore, we only considered agonist compounds from Assay #5.

Using our given information on chemical structure groups, we tabulated the groups to which the hits belonged. It was found that most hits came from group 2. Hence group 2 was pulled out from the data, and the same analysis was carried out to identify the hits from the other groups.

Finally, we compare hits derived from the above analysis with the hits found by the heuristic HTS analysis and obtained a 61.0% match. Comparison with the results of the statistical model and yielded a 74.5% match.

6.2 Logistic regression analysis

Given a series of n measurements on the compounds, "discriminants", one can ask which of them (if any) are good predictors of activity for the compounds, in the sense that large (small) values of the measurement tend to imply large (small) activity. The method of logistic regression is a useful tool for this purpose, and is described below.

Given a series of vectors $x^1, \ldots x^m \in R^n$, and associated real values $y^1, \ldots y^m$, it is a common problem to find a function that fits the dataset as best as possible in a given context, i.e. $f : R^m \to R$ such that $f(x^i) \approx y^i$ for every i. The logistic regression approach allows for functions of the form

$$f(x) = \frac{e^z}{1+e^z}, \quad z = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n,$$

where the choice of values β_i is determined with maximum likelihood methods. Note that the values of f are always in the interval $[0,1]$ so we interpret this function as the probability for a discrete event to hold for the point x.

In our context, each vector x^i represents a screened compound, or more precisely, an n-tuple of measurements performed on a compound. We hope that the measurements contain enough information about the compound for the application at hand, so that one can think of the compound as a point in Euclidean space.

The directions in which the values of the functions $f(x)$ and $z(x) = \beta_0+\beta_1 x_1+\ldots+\beta_n x_n$ increase are the same, and are described by the vector $\beta = (\beta_1, \ldots, \beta_n)$. If $\beta_i = 0$ for some i, we infer that the i^{th} measurement does not provide any predictive information about the value of f. On the contrary, if β_i is very large, we conclude that the i^{th} measurement might have a predictive value (given that large values of x_i imply large values of $f(x)$). In order to compare different values of β_i, $i \geq 1$, it is useful to normalize the measurements so that the sets $\{x_i^1 \ldots x_i^m\}$ have similar variance for every i.

We examined four descriptors: potency and efficacy for Proteins 1 and 2 (Assay #2). The choice of discriminants is biologically oriented so we might uncover more informative relationships than with typical chemical measurements, since it is suspected that these proteins are involved in B cell proliferation. We used only those compounds that were included in the LTS, and followed the ranking we had obtained to decide whether each compound should be considered active ($y = 1$) or inactive ($y = 0$). The Newton-Raphson method was used for finding the linear regression function $f(x)$, which can be interpreted as the probability for a compound to be active given the measurement vector x.

Results: Given the descriptors Potency_1, Efficacy_1, Potency_2, Efficacy_2, relating to Proteins 1 and 2 respectively, and using a total of 50 data points (37 active, 14 inactive), we obtained a logistic regression function with the parameters $\beta_0 = 1.4,\ \beta_1 = 1.18,\ \beta_2 = -0.27,\ \beta_3 = -0.24,\ \beta_4 = 2.12$. This indicates that Efficacy_2 is the best predictor of activity (out of the four) for a given compound. Surprisingly, the potency of the same protein does not seem to carry a predictive value. However, the potency of Protein 1 may be used as a (weak) predictor.

6.2.1 Testing hypotheses

In this section we make two short, tentative analyses of the relationships between Protein 1 and 2 secretion, activation of Receptors 1 and 2, and B-cell proliferation. It has been proposed that activation of Receptor 2 induces secretion of Protein 2, which in turn is associated with B cell proliferation. In what follows, we will restrict our attention to these two relationships.

Receptor 2 → Protein 2 Consider the 101 compounds that were tested both for receptor activation and protein potency/efficacy. We separated them into two groups, according to whether they were found to activate Receptor 2 or not. The protein secretion activity for the two groups can be described as follows:

	Protein 1 Efficacy	Protein 2 Efficacy
Receptor 2 activation	$\mu = 3.13,\ \text{MAD} = 0.212$	$\mu = 3.68,\ \text{MAD} = 0.36$
Receptor 2 non-activation	$\mu = 3.16,\ \text{MAD} = 0.24$	$\mu = 3.13,\ \text{MAD} = 0.44.$

We can see that the Protein 2 efficacy average is markedly higher for the group of Receptor 2-activating compounds than for the group of non-activating (or weakly activating) compounds. Indeed, the averages of 3.68 and 3.13 are significantly different, taking into account the MAD of each data set (recall that MAD, or median absolute difference, is a measure of the variance of the values, [2]). On the other hand, the Protein 1 efficacy average is quite similar for both groups (with respect to the corresponding MAD values) — a similar result holds for Protein 1 potency analysis. Thus, we have evidence of a correlation between Receptor 2 activation and Protein 2 secretion. In and of itself this doesn't imply a cause and effect relation, but it does support a scenario in which activating Receptor 2 triggers a signaling cascade resulting in protein production.

Note that in a similar way all binary relations between the given discriminants can be either established or refuted (to the extent that the data is sufficient and valid), and that both positive and negative influences can be distinguished.

Protein 2 → Cell Proliferation We considered the compounds that were screened for Protein 2 efficacy, and plotted them in Figure 9 using the corrected signal on the y-axis. The compounds that were deemed active using the techniques from the previous section are plotted in red.

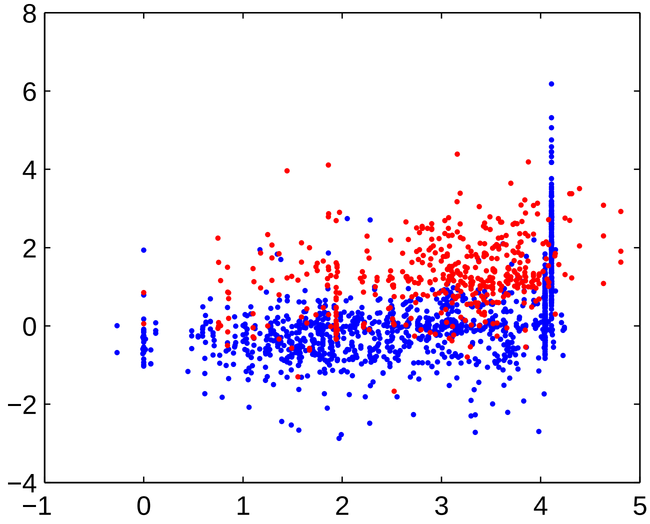

Figure 9: ?

It is not surprising that the red dots tend to have higher signal values, since their corresponding compounds were chosen on that basis. But it can also be seen that the red points tend to be clustered on the right side of the figure, suggesting a correlation between Protein 2 efficacy and B cell proliferation. Also, there doesn't seem to be another cluster of active points on the upper left corner which suggests that the only mechanism that the compounds have for inducing proliferation involves the secretion of Protein 2. This observation can be very useful when trying to establish a model for this process, but could not have been deduced only from a correlation analysis, since a small red cluster in the upper left side might not substantially diminish it. Finally, recall that active compounds were chosen solely on the basis of their signal, without involving any information on protein secretion.

7 Conclusions and future work

In the first half of this project we investigated in detail Assay #1, both HTS and LTS. LTS data was evaluated and compared to expert testimony to create a reference set of results. Next, different methods and models were developed for HTS and compared to LTS to determine the best method/model to identify HTS active compounds. The second half was devoted to the analysis of relationships between Assay #1 and other assays thought to be predictive of Assay #1.

If we had more samples per compound, we could use for example the approach described in [8]. In particular, let μ be the mean of the signal values, σ its standard deviation, n the number of samples and h_{ac} the threshold for compound activation (one possible choice is $h_{ac} = \mu + 3\sigma$). Then the probabilty of a compound being declared a hit could be calculated:

$$P(\text{hit}) = P(x > h_{ac}) = 1 - \Phi(\sqrt{n}\,\frac{h_{ac} - \mu}{\sigma}),$$

where Φ is the cumulative distribution function.

Another possibility of future work would be further investigation of common structural features of compounds. This is important because if only one compound of a family is considered to be a hit, it is a singlet and unlikely to be pursued because very little chemical optimization can be performed (toxicology for example).

If information about the HTS process were available, we would suggest modeling the HTS to decode the effect of environmental factors and other systematic noise. Also this would make it feasible to generate synthetic data to more easily validate algorithms for HTS data analysis.

In this project we focused on heuristics and abstract statistical models—that is, methods based on data mining. Those models were suited to the problem posed and data available. We believe that an better understanding of the environmental and biological variables in HTS would allow us to construct a descriptive model, which would provide more insight into the HTS process.

8 Acknowledgements

We would like to thank the Institute for Mathematics and its Applications and the University of Minnesota for hosting the Workshop for Graduate Students: Mathematical Modeling in Industry. We would also like to thank 3M Corporation and our mentor Dr. Ann DeWitt for their support and guidance, and Professor Douglas Arnold for suggesting statistical modeling, [7].

References

[1] G.A. Bishop, L.M. Ramirez, M. Baccam, L.K. Busch, L.K. Pederson, and M.A. Tomai. The immune response modifier resiquimod mimics cd40-induced b cell activation. *Cellular Immunology*, 208:9–17, 2001.

[2] C. Brideau, B. Gunter, B. Pikounis, and A. Liaw. Improved statistical methods for hit selection in high-throughput screening. *The Society for Biomolecular Screening*, 8(6):634–647, 2003.

[3] P. Gedeck and P. Willet. Visual and computational analysis of structure-activity relationships in high-throughput screening data. *Current Opinion in Chemical Biology*, 5:389–395, 2001.

[4] J. Hallborn and R. Carlsson. Automated screening procedure for high-throughput generation of antibody fragments. *BioInvent Therapeutic AB*, 33:530–537, 2002.

[5] S. Heyse. Comprehensive anlysis of high-throughput screening data. Technical report, GeneData AG, http://www.genedata.com, 2001.

[6] T. Ideker and D. Lauffenburger. Building with a scaffold: Emerging strategies for high-to-low-level cellular modeling. *TRENDS in Biotechnology*, 21(6):255–262, 2003.

[7] A. J. Lichtman and V. I. Keilis-Borok. Pattern recognition applied to presidential elections in the United States, 1860–1980: Role of integral social, economic, and political traits. *Proceedings of the National Academy of Sciences*, 78(11):7230–7234, 1981.

[8] G.S. Sittampalam, P.W. Iversen, J.A. Boadt, S.D. Kahl, J.M. Zock S. Bright, W.P. Janzen, and M.D. Lister. Design of signal windows in high throughput screening assays for drug discovery. *Journal of Biomolecular Screening*, 2(3):156–169, 1997.

[9] Y. Wang, H.A. Chipman, and W.J. Welch. Mining nuggets of activity in high dimensional space from high throughput screening data. Technical report, IIQP, 2002.

[10] J. Zhang, T.D.Y. Chung, and K.R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening*, 4(2):67–73, 1999.