BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2271–2278
doi:10.1093/bioinformatics/btr369

Data and text mining

Advance Access publication June 17, 2011

Enhancing the rate of scaffold discovery with diversity-oriented prioritization

S. Joshua Swamidass[1,2,*], Bradley T. Calhoun[1], Joshua A. Bittker[2], Nicole E. Bodycombe[2] and Paul A. Clemons[2]

[1]Division of Laboratory and Genomic Medicine, Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO and [2]Chemical Biology/Novel Therapeutics, The Broad Institute of Harvard and MIT, Cambridge, MA, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In high-throughput screens (HTS) of small molecules for activity in an *in vitro* assay, it is common to search for active scaffolds, with at least one example successfully confirmed as an active. The number of active scaffolds better reflects the success of the screen than the number of active molecules. Many existing algorithms for deciding which hits should be sent for confirmatory testing neglect this concern.

Results: We derived a new extension of a recently proposed economic framework, diversity-oriented prioritization (DOP), that aims—by changing which hits are sent for confirmatory testing—to maximize the number of scaffolds with at least one confirmed active. In both retrospective and prospective experiments, DOP accurately predicted the number of scaffold discoveries in a batch of confirmatory experiments, improved the rate of scaffold discovery by 8–17%, and was surprisingly robust to the size of the confirmatory test batches. As an extension of our previously reported economic framework, DOP can be used to decide the optimal number of hits to send for confirmatory testing by iteratively computing the cost of discovering an additional scaffold, the marginal cost of discovery.

Contact: swamidass@wustl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 18, 2011; revised on June 10, 2011; accepted on June 14, 2011

1 INTRODUCTION

All screeners must decide which initial positives ('hits') from a high-throughput screen (HTS) to submit for confirmatory experiments (Makarenkov *et al.*, 2007; Nicholls, 2008; Rocke, 2004; Storey *et al.*, 2007). In HTS of small molecules for biological activity, several hits are often experimentally confirmed by ensuring they exhibit the characteristic dose–response behavior common to true actives. Sending the wrong molecules for confirmatory testing wastes resources, reducing the amount of useful information in screening results; this is especially true in the context of small molecule screens which are often very noisy.

*To whom correspondence should be addressed.

Most commonly used hit selection methods focus on maximizing the number of successfully confirmed molecules in subsequent confirmatory experiments. These methods include correcting time- and well position-dependent systematic error in measurements (Makarenkov *et al.*, 2007), exploiting chemical information (Glick *et al.*, 2004, 2006; Posner *et al.*, 2009), better normalizing screening data (Zhang *et al.*, 2005), choosing better experimental controls (Seiler *et al.*, 2008) or bypassing hit selection entirely by testing all molecules in dose–response experiments (Inglese *et al.*, 2006). These methods are effective, increasing the number of successful confirmations, but largely ignore the fact that screeners are often looking to maximize the total number of clusters–groups of molecules with similar structure–containing examples with confirmed activity (Clark and Webster-Clark, 2008).

Some methods use molecular clustering to improve the design of HTS experiments. For example, molecular clusters have been used to reduce the total number of molecules in the primary screen by about two-thirds (Karnachi and Brown, 2004). Similarly, other studies pick molecules for follow-up using statistical tests on the data from a single-dose screen to find clusters of active molecules (Varin *et al.*, 2010; Yan *et al.*, 2005). These methods favor clusters that contain several active molecules, and attempt to send all these active molecules for confirmatory testing. As intended, these methods successfully increase the number of active molecules identified in confirmatory testing. In contrast, diversity-oriented prioritization (DOP) aims to maximize the diversity of confirmed actives by maximizing the number of clusters with at least one successfully confirmed active. Rather than picking groups of similar molecules, DOP picks molecules from as many different groups as possible given the cost constraint. This aim is motivated by the fact that screeners often cluster confirmed active molecules to pick series of molecules to optimize, and that the number of series with at least one confirmed active is a reasonable way of measuring the information obtained from a screen (Clark and Webster-Clark, 2008).

The DOP method extends a recently described economic framework for interpreting HTS data, initially introduced to decide how many hits to send for confirmatory testing (Swamidass *et al.*, 2010). This framework is used to iteratively choose each batch of hits to be sent for confirmatory testing so as to maximize the expected surplus of the batch. The expected surplus is computed using three mathematical models: utility, cost and predictive. The utility model specifies the preferences of the screener, the cost model tracks

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

the cost of running a confirmatory experiment and the predictive model guesses the outcome of future confirmatory experiments. DOP extends this framework by introducing a new utility model, from which we derive a new method of prioritizing hits.

We validated the DOP methodology and the algorithm that implements it with both retrospective and prospective experiments. These experiments demonstrate that DOP can substantially increase the number of active scaffolds discovered in a HTS experiment.

2 DATA

The technical details of the assay and subsequent analysis can be found in PubChem (PubChem AIDs 1832 and 1833). For ~300 000 small molecules screened in duplicate, we defined activity as the mean of final, corrected percent inhibition. After molecules with autofluorescence and those without additional material in stock were filtered out, 1322 with activity greater than 25% were labeled 'hits' and tested for dose–response behavior in the first batch. Of these tested molecules, 839 yielded data consistent with inhibitory activity. Each hit was considered an 'active' if the effective concentration at half maximal activity (EC50) was less than or equal to 20 μM. Using this criterion, we determined 410 molecules to be active.

3 METHODS

3.1 Scaffold clusters

There are two common strategies used to cluster small molecules: similarity-based and scaffold-based clusterings (Bemis and Murcko, 1996; Butina, 1999; Clark and Labute, 2008; Downs *et al.*, 1994; Schuffenhauer *et al.*, 2007; Shemetulskis *et al.*, 1995; Willett, 2006). Similarity-based clustering groups structurally similar molecules—as measured by fingerprint similarity—together. Within each cluster, molecules' structures are very close, but it may not be possible to align molecules because there may not be substructures common to all the molecules in the cluster. In contrast, scaffold-based clustering groups molecules into clusters with a well-defined common substructure. Therefore, within each scaffold cluster, molecules are easily aligned (Clark and Labute, 2008).

Often, HTS campaigns aim to identify as many new scaffolds as possible. Sometimes intellectual property concerns dictate both avoiding particular scaffolds and defining discoveries using the scaffold concept. Scaffolds are often the starting points from which lead-refinement proceeds (Clark and Webster-Clark, 2008; Good and Oprea, 2008; Schuffenhauer *et al.*, 2007). Therefore, we focused on scaffold-based clustering. Nonetheless, our methods can be easily adapted to similarity-based clustering.

We computed scaffolds from the structure of each molecule using the molecular framework algorithm described by Bemis and Murcko (1996): contiguous ring systems and the chains that link two or more rings together. Molecular frameworks are only an approximation of a medicinal chemists subjective concept of a scaffold (Clark and Labute, 2008; Schuffenhauer *et al.*, 2007). Nonetheless, frameworks are commonly used in chemical informatics because they are clearly defined and easy to compute. Although we define scaffolds as molecular frameworks, DOP is compatible with more sophisticated scaffold detection algorithms; all it requires is that molecules are grouped appropriately.

In order to ensure our findings were not overly dependent on the choice of scaffold definition, all experiments in this study were replicated using a modification of each scaffold that replaces every atom in the scaffold with a carbon. The results of this variation are not presented because they are not notably different. This observation suggests that similar results would also be observed with other scaffold-detection algorithms, though we have not directly verified this.

3.2 Utility model

The preferences of screeners are difficult to assess and often inconsistent between different experts (Lajiness *et al.*, 2004). Nonetheless, some parts of their preferences can be modeled. In this study, a scaffold was considered to be 'active' if at least one example of the scaffold was confirmed as active. Consistent with prior work (Clark and Webster-Clark, 2008), one unit of discovery was defined as a single active scaffold. Of course, more robust definitions of an active scaffold are possible—for example, defining scaffolds active if they have two or three confirmed active examples—however, these definitions require more complicated algorithms to implement and will be elaborated in future work. The utility model $U(D)$ was defined as a function of the total discovery so far, D: the number of scaffolds with at least one example confirmed active, corresponding with maximizing the number of scaffolds identified, giving chemists maximally diverse candidate starting points for follow-up chemistry.

3.3 Cost model

The cost model used is relevant to the implementation of DOP in two ways. First, in some scenarios, the cost of acquiring different molecules varies. In the context of HTS, however, molecules under consideration are usually equally accessible. Therefore, we assumed that it costs the same amount to send each molecule for confirmatory testing. Second, there are both large fixed and smaller variable costs associated with sending molecules for confirmatory tests. Under these circumstances, confirmatory tests are most efficiently performed in large batches, just as is done in practice.

3.4 Predictive model

We considered two predictive models: a logistic regressor (LR) (Dreiseitl and Ohno-Machado, 2002) and a neural network with a single hidden node (NN1) (Baldi and Brunak, 2001). Both are structured to use the screen activity as the single independent variable and the result of the associated confirmatory experiment as the single dependent variable. Networks with more hidden layers work as well, but do not yield substantially better results. Both the LR and NN1 models were trained using gradient descent on the cross-entropy error using the monotonic prior defined in the Appendix A (with $k=2$ and $\theta=0.5$) along with a Gaussian prior on weights not addressed by the monotonic prior (Baldi and Brunak, 2001). In general, such a protocol yields models whose outputs are interpretable as a probabilities,

$$P(x)\begin{cases}1 & \text{if molecule } x \text{ is confirmed active}\\ 0 & \text{if molecule } x \text{ is confirmed inactive,}\\ z_x & \text{if molecule } x \text{ has not been tested}\end{cases} \quad (1)$$

where z_x is the output of the predictive model on the molecule x. There is little distinction between LR and NN1 in practice and, in fact, virtually any probabilistic method can be used by the DOP method.

3.5 Prioritization algorithm

The economic framework prescribes choosing the next batch to maximize the expected surplus (ES) after the next batch is screened,

$$E\big[U(D'+D)\big]-(C+C') \quad (2)$$

where D' is the number of scaffolds in the next batch, D is the number of scaffolds discovered so far, C is the cost expended confirming molecules so far and C' is the cost of screening the next batch. Removing the cost terms, which are constant across all molecules, shows that maximizing the ES is equivalent to maximizing expected utility (EU),

$$E\big[U(D'+D)\big]. \quad (3)$$

Furthermore, because $U(\cdot)$ is likely to be approximately linear over the narrow distribution of D', this equation is well approximated by,

$$U\big(E[D']+D\big), \quad (4)$$

a well-studied approximation from the economics literature (Levy and Markowitz, 1979; Schoemaker, 1982). Because $U(\cdot)$ is monotonically

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

increasing, maximizing the EU—and also maximizing ES—is approximately equivalent to maximizing the expected marginal discovery (EMD)

$$E[D'], \tag{5}$$

and therefore we propose prioritizing molecules by choosing the next batch of molecules so as to maximize the EMD of the batch.

A different protocol is required to select the molecules in the first batch because it is impossible to compute the EMD without knowing the results of at least some confirmatory experiments. For the first batch, we chose molecules with the top activity in the screen while at the same time assuring that no more than one example of each scaffold was selected.

3.6 Computing EMD

The algorithm for computing the EMD requires the output from the predictive model and assumes that each probability of successful confirmation is independent of the others, and that within a scaffold group each molecule is screened in the order of decreasing probability of activity. With these assumptions, the EMD of the x-th molecule in the scaffold is

$$\mathrm{EMD}(x) = P(x\text{-th is first active}), \tag{6}$$

where $P(x\text{-th is first active})$ is the probability molecule x is the first confirmed active within its scaffold group, and the total EMD of the next batch $E[D']$ is computed as the sum of the EMDs associated with each molecule in the batch.

To do this, we used an algorithm to compute $P(x\text{-th is first active})$. For each group of molecules with a common scaffold, we defined the probability that each molecule is active using Equation (1). The molecules in the group were sorted in decreasing order of probability so that $\{P(1) \geq P(2) \geq P(3) \geq \ldots\}$. Assuming the untested molecules must be prioritized in this order, the probability that the j-th molecule in this list is the first active is

$$\mathrm{EMD}(j) = P(j \text{ is first active}) = P(j)\prod_{k=1}^{j-1}\left[1 - P(k)\right], \tag{7}$$

where the probability that molecule j is active, $P(j)$, is multiplied by the probability that all prior molecules in the scaffold group are inactive, $\prod_{k=1}^{j-1}[1-P(k)]$.

The EMD of the next batch was thus maximized by choosing the untested molecules with highest EMD. These EMD's remain ordered in decreasing order so that $\{\mathrm{EMD}(1) \geq \mathrm{EMD}(2) \geq \mathrm{EMD}(3) \geq \ldots\}$. Within a scaffold group, therefore, untested molecules were prioritized in the same order as their initial HTS activity. Depending on how many total molecules are to be tested, the algorithm will usually pick one molecule from each scaffold but will occasionally choose more than one. With respect to molecules from other scaffold groups, however, their order may be shuffled: exactly the behavior we seek.

4 RESULTS

4.1 Scaffold distribution

The skewed distribution of scaffolds in HTS data motivated our approach. Ignoring this distribution, resources would be wasted trying to confirm examples of scaffolds that have already been discovered; in other words, there is significant redundancy both in HTS libraries and the hits from screens of these libraries. We considered three sets of molecules: the full set of molecules from the screen (the Library), the first batch sent for dose–response confirmation (the Dosed) and the molecules subsequently confirmed as active (the Active). There were 301 617, 1322 and 410 molecules, and 84 440, 1043 and 331 scaffolds, respectively, in each of these sets. Each scaffold is represented by, on average, 3.57, 1.27 and 1.24 examples. The molecules are not distributed evenly among scaffolds; a few scaffolds are disproportionately frequent, but > 50% of the Dosed set is composed of scaffolds with only one example (Table 1). These results reflect the prior observation that molecules follow a power-law distribution when clustered (Benz *et al.*, 2008). Similar distributions would be expected if the structures had been clustered by almost any other clustering algorithm.

Table 1. Scaffold distributions

Rank	Number of examples			Percentage of data		
	Library	Dosed	Active	Library	Dosed	Active
1	7215	19	6	2.3	1.4	1.5
2	1104	10	5	0.4	0.8	1.2
3	886	8	5	0.3	0.6	1.2
4	797	8	5	0.3	0.6	1.2
5	704	7	4	0.2	0.5	1.0
Frequency	**Number of scaffolds**			**Percentage of data**		
	Library	Dosed	Active	Library	Dosed	Active
1	42889	887	285	14.2	67.1	69.5
2	14599	101	26	9.7	15.3	12.7
3	8065	30	12	8.0	6.8	8.8
4	5132	10	4	6.8	3.0	3.9
5	3382	9	3	5.6	3.4	3.7

The 'Library' is the approximately 300 000 molecules screened in the initial HTS assay, 'Dosed' are the molecules that were sent for dose-response confirmation in the first batch, and 'Active' are those molecules subsequently confirmed as active. For each set of molecules, the top table displays the frequency of the top five most common scaffolds and the percentage of the total dataset that each scaffold group represents. The bottom table displays the number of scaffolds with exactly 1, 2, 3, 4 or 5 examples in the data (frequency) and the percentage of data that these scaffold groups represent.

The scaffold distribution sets an upper bound on DOP's efficiency. Of the 1322 molecules sent for dose–response experiments, only 79% could be, in a best-case scenario, the first example of their scaffold group. Therefore, at most we could expect a 21% reduction in the number of confirmatory experiments required to discover a fixed number of scaffolds. This estimation is based on one dataset and would need to be revised upwards or downwards with other datasets. Furthermore, this estimate only considers the proportion of singletons; more accurate predictions could be constructed. Nonetheless, it provides a useful theoretical baseline against which to gauge DOP's empirical performance.

4.2 Predicting yield

One test of the DOP algorithm is to assess whether it can accurately predict yield—the number of scaffold discoveries in a batch of experiments—using Equations (6) and (7) in conjunction with either of the two predictive models, LR and NN1.

In this experiment, 1322 molecules with known dose–response outcomes were ordered by their initial HTS activity. They were then divided into plates of 30 molecules each. The yield of each plate was predicted by training a predictive model (LR or NN1) on the outcome of all prior confirmatory tests, as described in the Section 3. The predicted probabilities of activity were then used in conjunction with Equations (6) and (7) to yield a final prediction. The predicted yield is close to the empirical number of discoveries (Fig. 1).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

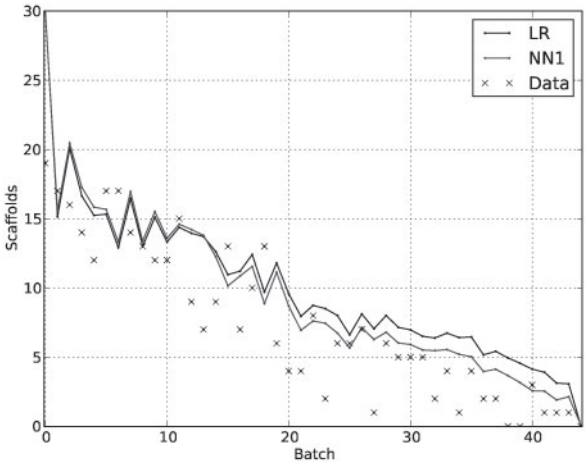

Fig. 1. Predicted confirmation rate of scaffolds. The per-batch predicted discoveries of active scaffolds (LR and NN1) along with observed discoveries (labeled 'Data').

This experiment demonstrates that the DOP algorithm can predict the number of scaffold discoveries in a plate. There is some systematic inaccuracy in these predictions; both LR or NN1 seem to overestimate the number of discoveries by a small amount. It is possible that this systematic error is due to dependencies in the data ignored by our model: for example, successfully confirmed actives are likely to share a common scaffold. Nonetheless, the predictions are close to the observed yield.

4.3 Reordering hits

Another test of DOP is to verify that it modifies the order in which molecules are sent for confirmatory testing. Furthermore, because confirmatory experiments are usually batched in large groups, it is important to study how the DOP rankings change as the batch sizes are varied. In this experiment, we compared the order of molecules ranked by initial HTS activity with the ordering generated by the DOP algorithm. For comparison, the DOP algorithm was run six times using two different predictive models and three different batch sizes: 1 (unbatched), 30 and 300.

This experiment yields several important observations. First, both NN1 and LR yielded almost identical rankings ($R^2 = 0.999$). This is an important result that suggests that DOP is robust to the predictive model: subtle differences in the predictive model may not dramatically affect how molecules are ordered. Second, DOP may be surprisingly robust to batch size. Using a batch size of 30 was virtually identical to the unbatched DOP (Fig. 2). There were more noticeable differences with unbatched DOP and DOP using a batch size of 300. Nonetheless, using DOP with a batch size of 300 yielded rankings much closer to the unbatched DOP rankings than to the original HTS rankings. Finally, in all cases, $\sim$200 molecules were not ranked because their EMDs were equal to zero before they were prioritized. This is exactly the desired behavior; testing these 200 molecules would be redundant because they have scaffolds that had already been confirmed active.

4.4 Increasing scaffold discovery rate

The most important *in silico* test of DOP is to verify that it increases the rate of scaffold discovery. While it is clear that DOP reorders molecules relative to the HTS activity, the rate at which scaffolds are discovered must increase in order to conclude that DOP is preferable to ordering molecules by HTS activity.

In this experiment, molecules were prioritized by DOP in batches of 30. Compared with prioritizing by HTS activity, the total number of scaffolds discovered was plotted against the total number of confirmatory experiments (Fig. 3). DOP shifts the curve upward, indicating that for any specific number of confirmatory experiments, more scaffolds were discovered using DOP. On average, DOP discovered 1.18 more scaffolds per batch than HTS activity prioritization. DOP increased the scaffold discovery rate.

4.5 Effect of batch size

For DOP to be useful in practice, it must be robust to large batch sizes. We might expect a trade-off between batch size and efficiency; larger batch sizes might decrease DOP efficiency due to inaccuracies in the predictive model and uncertain outcomes in confirmatory experiments. However, the ranking data suggest that DOP may be robust to batch size. We performed a more comprehensive test to resolve these conflicting expectations.

In this experiment, DOP was run with several different batch sizes using both predictive models. Compared with ordering molecules by HTS activity, the number of confirmatory experiments required to discover 50, 75 and 100% of the scaffolds was reduced by 8–17% (Table 2). Surprisingly, these numbers were largely consistent (and often identical) across all batch sizes and predictive models. Batch size did not appreciably affect the rate at which scaffolds are discovered; DOP is robust to batch size.

4.6 Prospective validation

Although these retrospective experiments are promising, the most important test of DOP is a prospective experiment. In this experiment, we used a batch size of 500 molecules. DOP, using both LR and NN1, was used to pick the next 500 molecules to test. These two lists of candidates were compared with the next batch of 500 molecules selected by HTS activity. The yield of all three lists was predicted and as many molecules as were available were obtained and sent for confirmatory testing. In this case, of the 500 compounds suggested, 479 from the LR and NN1 batches and 477 from the HTS batch were sent for testing.

Even without the results of the confirmatory tests, this experiment reinforced several results from the retrospective experiments. First, the batches suggested by LR and NN1 were almost identical, with only 10 molecules different, corresponding with the observation that LR and NN1 yielded almost identical rankings in the retrospective experiments. Second, both DOP batches were substantially different than the HTS batch, with 105 molecules different in the LR list and 95 different in the NN1 list, again corresponding with similar observations from the retrospective experiments. More importantly, LR predicted a 12.5% increase in scaffold discovery (126 compared with 112). Likewise, NN1 predicted a similar 8.7% increase in the rate of scaffold discovery (113 compared with 104).

The results of the confirmatory testing also correspond with the results from the retrospective experiments. First, and most importantly, more scaffolds were discovered in the DOP batches; 170 and 166 scaffolds were discovered in the LR and NN1 batches, compared with 153 scaffolds in the HTS batch. Finally, the predicted yield was close to, but still underestimated, the actual

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

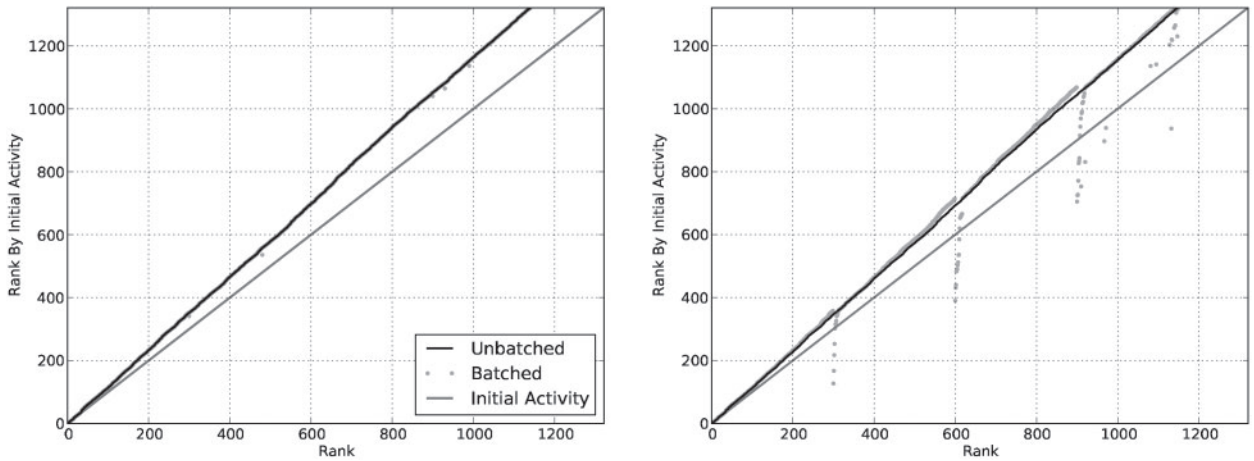

Fig. 2. Changes in rank. Comparison of ranking by DOP selection to ranking by HTS results. 'Initial Activity' is the order of molecules ranked by HTS results. 'Unbatched' shows the rankings of DOP computed in a molecule-by-molecule protocol (e.g. batch size of 1), while 'Batched' corresponds to DOP computed using more realistic batch sizes of (left) 30 and (right) 300 molecules. Identical results are produced by both NN1 and LR.

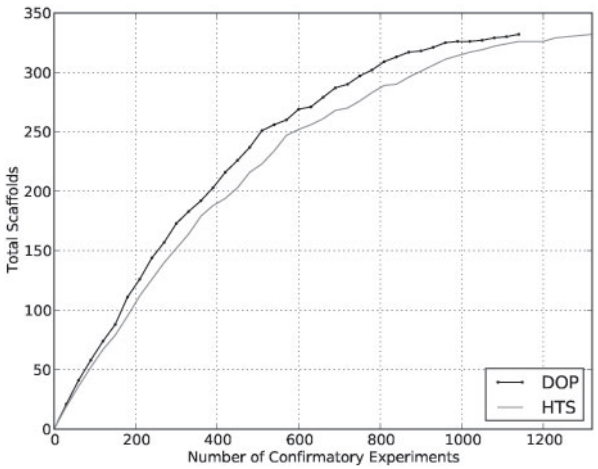

Fig. 3. Improvement in scaffold discovery rate. 'DOP' indicates the scaffolds discovered using DOP, done in batches of 30 molecules. The order of selection for confirmatory experiment given by NN1 (same as order of selection by LR). 'HTS' is the scaffolds discovered when tested in the order of their HTS results.

number of scaffolds discovered: LR predicted DOP to discover 126 (actual 170) and HTS to discover 112 (actual 153). NN1 predicted DOP to discover 113 (actual 166) and HTS to discover 104. The predictions are underestimations of the actual yields. Predicted increase in scaffold discovery were more accurate. LR predicted 12.5% more scaffolds discovered in DOP than HTS (actual 11.1%). NN1 predicted 8.7% more (actual 8.5%).

5 DISCUSSION

In both retrospective and prospective experiments, DOP increased the rate of scaffold discovery from an HTS experiment. The key result of this study is that screeners preferences, when more accurately modeled, can be applied to change the order in which molecules are tested so as to increase the rate at which active scaffolds are discovered.

Table 2. Batch size and discovery rate

Batch size	75% sensitivity		100% sensitivity	
	LR (%)	NN1 (%)	LR (%)	NN1 (%)
1	502 (14)	501 (14)	1124 (14)	1124 (14)
30	501 (14)	501 (14)	1124 (14)	1124 (14)
100	499 (14)	499 (14)	1124 (14)	1124 (14)
200	498 (15)	498 (15)	1121 (15)	1119 (15)
300	504 (14)	504 (14)	1117 (15)	1119 (15)
400	498 (15)	498 (15)	1132 (14)	1136 (13)
500	514 (12)	514 (12)	1123 (14)	1126 (14)

Confirmatory tests run on the same set of molecules, sorted by DOP with different batch sizes. The table indicates the number of confirmatory experiments required before some percentage (75 or 100%) of the total active scaffolds in the set were discovered. Percentages in parenthesis indicate improvement with respect to confirmatory tests run in order of HTS results.

Several additional issues arise when using DOP in practice. First, there will be several molecules that are very likely active but will not be sent for confirmatory testing. These molecules are de-prioritized because they belong to scaffold groups with at least one example already confirmed active. This is not the behavior some might expect from a prioritization algorithm. One extension to address this concern would be to report the probability of activity—the output of LR or NN1—for all the molecules from scaffold groups with at least one confirmed active. A succinct way to summarize this information is to report the total number of actives and inactives expected if all the molecules in the group had been tested (Fig. 4).

DOP could be used as the first stage of a two-stage prioritization protocol, where the first few batches are selected to maximize scaffold discovery. In the second stage, with the screener's input, a number of scaffolds could then be selected for follow-up experiments based on their structures, the initial results from the confirmatory experiments and the predicted number of actives and inactives associated with them. Additional examples of the selected scaffold would then be sent for confirmatory testing to build enough

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

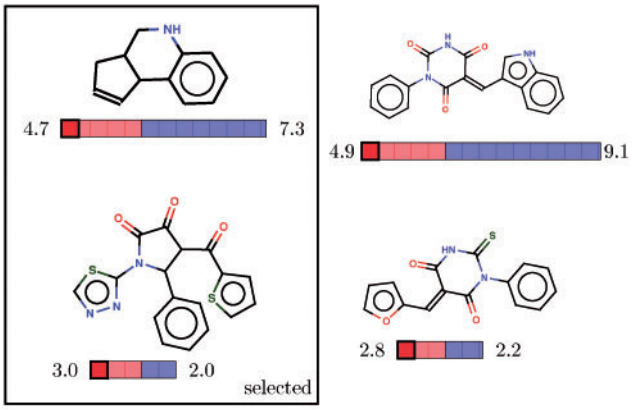

Fig. 4. This display is sufficiently informative to enabled chemists to prioritize singletons for follow-up after a screen. All four scaffolds have only one confirmed active (dark red block), with the predicted number of actives (light red bars and left numbers) and inactives (blue bars and right numbers) displayed below each structure. These predictions are fractional, reflecting the expected number of actives and inactives for each scaffold. The top-left scaffold's structure and predicted number of actives were both favorable to a particular medicinal chemist. The bottom-left scaffold's questionable structure was outweighed by the number of predicted actives. The right scaffolds' structures were so unfavorable that they were considered undesirable regardless of the number of actives confirmed; the chemist believed that they were more difficult to chemically modify. Of course, other medicinal chemists might rank these structures differently based on their experience and intentions.

evidence to establish the scaffold as a true active. Such a two-step protocol, first maximizing scaffold discovery then confirming more examples of interesting scaffolds, provides an example of how our methods could be modified to better model screeners' preferences. Better utility functions might seek to maximize the number of scaffolds with a clique of at least two or three confirmed actives. Clique-oriented prioritization (COP), however, requires a more complicated algorithm that will be presented in future work.

Importantly, DOP can and should be used simultaneously with other HTS analysis algorithms. For example, other prioritization methods designed to reduce false positives—by using better controls, chemical information, or other strategies—are all compatible with DOP; the priorities generated by these methods can either be substituted for the HTS activity or be presented as an additional independent variable to the predictive model.

Furthermore, DOP is a direct extension of a previously defined economic framework and can, therefore, be used to compute the marginal cost of discovery (MCD) (Swamidass *et al.*, 2010). The MCD is the cost required to discover one more scaffold, and yields an optimal strategy for deciding how many and which molecules should be sent for confirmatory testing. The MCD is effectively the price of the next active, and the screener should keep screening molecules until the MCD rises too high and the utility of the next scaffold is not worth the cost of finding it.

DOP is not without limitations. Most importantly, DOP relies on a definition of an active scaffold that may not be optimal. Furthermore, DOP requires that the results from enough confirmatory experiments are known for the predictive model to be trained. In contrast, most other HTS hit selection methods are applied to primary HTS data without knowing anything about the outcome of any confirmatory experiments (Karnachi and Brown, 2004; Varin *et al.*, 2010; Yan *et al.*, 2005). This limitation essentially requires HTS to proceed iteratively, with at least two batches. Future work will include methods of circumventing this limitation by using predictive models that do not require confirmatory experiments to be parameterized.

Also, as presented, DOP assumes that there are no errors in the confirmatory experiment, that each molecule's potency is measured accurately in the dose–response experiment. However, just like the primary screen, there can be substantial noise in the confirmatory experiment (Eastwood *et al.*, 2006). One method of accounting for errors in the confirmatory experiment is by labeling molecules by their *probability* of having a satisfactory potency when taking the noise of the assay into account. This strategy would, likely, improve the quality of the predictive models and more comprehensively address the problem of noise in HTS projects. A detailed description and evaluation of this approach will be considered in future work.

6 CONCLUSION

When screeners' preferences are modeled, they can be applied to change which hits are selected from a high-throughput screen and sent for confirmatory testing. For example, DOP relies on the observation that screeners are looking for active scaffolds more than they are looking for active molecules. In this study, both retrospective and prospective experiments demonstrate that DOP can increase the rate of scaffold discovery from 8% to 17%. DOP is robust to batch size, and can be applied even when using very large batch sizes common in HTS experiments. More accurate models of screeners' preferences may prove even more useful.

ACKNOWLEDGEMENT

Marvin was used to generate the chemical structures in Figure 4; Marvin 5.3.5, 2010, ChemAxon (http://www.chemaxon.com).

Funding: We thank the Physician Scientist Training Program of the Washington University Pathology Department for supporting S.J.S. and B.T.C.; J.A.B., N.E.B. and P.A.C. are supported in part by the NIH Molecular Libraries Network (U54-HG005032) and the NIGMS-sponsored Center of Excellence in Chemical Methodology and Library Development (P50-GM069721).

Conflict of Interest: none declared.

REFERENCES

Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*. The MIT Press, Cambridge, MA.

Bemis,G.W. and Murcko,M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, **39**, 2887–2893.

Benz,R. *et al.* (2008) Discovery of power-laws in chemical space. *J. Chem. Iinform. Model.*, **48**, 1138.

Butina,D. (1999) Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inform. Comput. Sci.*, **39**, 747–750.

Clark,A. and Labute,P. (2008) Detection and assignment of common scaffolds in project databases of lead molecules. *J. Med. Chem.*, **52**, 469–483.

Clark,R. and Webster-Clark,D. (2008) Managing bias in ROC curves. *J. Comput. Aided Mol. Des.*, **22**, 141–146.

Downs,G. *et al.* (1994) Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inform. Comput. Sci.*, **34**, 1094–1102.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Dreiseitl,S. and Ohno-Machado,L. (2002) Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.*, **35**, 352–359.
Eastwood,B. *et al*. (2006) The minimum significant ratio: a statistical parameter to characterize the reproducibility of potency estimates from concentration-response assays and estimation by replicate-experiment studies. *J. Biomol. Screen.*, **11**, 253.
Glick,M. *et al*. (2004) Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screen.*, **9**, 32.
Glick,M. *et al*. (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inform. Model.*, **46**, 193–200.
Good,A. and Oprea,T. (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput. Aided Mol. Des.*, **22**, 169–178.
Inglese,J. *et al*. (2006) Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl Acad. Sci. USA*, **103**, 11473.
Karnachi,P. and Brown,F. (2004) Practical approaches to efficient screening: information-rich screening protocol. *J. Biomol. Screen.*, **9**, 678.
Lajiness,M. *et al*. (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.*, **47**, 4891–4896.
Levy,H. and Markowitz,H. (1979) Approximating expected utility by a function of mean and variance. *Am. Econ. Rev.*, **69**, 308–317.
Makarenkov,V. *et al*. (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics*, **23**, 1648.
Nicholls,A. (2008) What do we know and when do we know it? *J. Comput. Aided Mol. Des.*, **22**, 239–255.
Posner,B.A. *et al*. (2009) Enhanced HTS hit selection via a local hit rate analysis. *J. Chem. Inform. Model.*, **49**, 2202–2210.
Rocke,D. (2004) Design and analysis of experiments with high throughput biological assay data. *Seminars in Cell and Developmental Biol.*, **15**, 703–713.
Schoemaker,P. (1982) The expected utility model: its variants, purposes, evidence and limitations. *J. Econ. Lit.*, pp. 529–563.
Schuffenhauer,A. *et al*. (2007) The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inform. Model.*, **47**, 47–58.
Seiler,K. *et al*. (2008). ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351.
Shemetulskis,N. *et al*. (1995) Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput. Aided Mol. Des.*, **9**, 407–416.
Storey,J. *et al*. (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8**, 414.
Swamidass,S.J. *et al*. (2010) An economic framework to prioritize confirmatory tests after a high-throughput screen. *J. Biomol. Screen.*, **15**, 680–686.
Varin,T. *et al*. (2010) Compound set enrichment: a novel approach to analysis of primary HTS data. *J. Chem. Inform. Model.*, **50**, 277–279.
Willett,P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.
Yan,S. *et al*. (2005) Novel statistical approach for primary high-throughput screening hit selection. *J. Chem. Inform. Model.*, **45**, 1784–1790.
Zhang,J.H. *et al*. (2005) Probing the primary screening efficiency by multiple replicate testing: a quantitative analysis of hit confirmation and false screening results of a biochemical assay. *J. Biomol. Screen*., **10**, 695–704.

APPENDIX A

A.1 Monotonic models

Models like LR and NN1 estimate the correct weights by using gradient descent to maximize the log likelihood of the model with respect to the training data:

$$L=\sum_i\Big(t_i\log z_i(\overline{w})+(1-t_i)\log\big[1-z_i(\overline{w})\big]\Big), \tag{8}$$

where i ranges over the training data, $t_i\in\{1,0\}$ is the outcome of the confirmation experiment on the i-th training example, z_i is the output of the model on the i-th training example and $\overline{w}$ is the weight vector whose elements are chosen to maximize the log likelihood L.

Sometimes this process yields weights that are inaccurately estimated. Usually, the inaccuracy only subtly affects the model's yield predictions and the order in which molecules are prioritized. Occasionally, when trained on small amounts of data, this inaccuracy can be more dramatic. The worst failure occurs when a model inaccurately learns that the molecules with the worst activity in the initial screen are the most likely to be active in the confirmatory experiment. This, in turn, prioritizes the most clearly inactive molecules from the initial screen above all others.

This failure can be prevented by ensuring that the model is positive monotonic: that its output always increases as its input—the activity in the initial screen—increases. We accomplish this by adding a gamma prior to components of the weight vector. In the case of LR, the prior is placed on the weight multiplied by the input data, but not on the weight added to the product of this multiplication. In the case of NN1, the prior is placed on the weights that multiply either input or hidden nodes, but not on any of the threshold weights. Gamma priors on these weights ensure that trained models will always be positively monotonic with respect to their inputs no matter what the training data.

The gamma prior is implemented by defining a probability distribution over select weights in the model,

$$P(\overline{w})=\prod_j w_j^{k-1}\frac{e^{-w_j/\theta}}{\theta^k\Gamma(k)}, \tag{9}$$

where j ranges over the components of $\overline{w}$ to which the gamma priors are applied. k and θ are the position and shape parameters of the distribution, and $\Gamma(\cdot)$ is the gamma function. This amounts to changing the objective of the training algorithm to maximize likelihood,

$$L=\sum_j\Big[(k-1)\log w_j-\frac{w_j}{\theta}\Big]+\sum_i\Big(t_i\log z_i(\overline{w})+(1-t_i)\log\big[1-z_i(\overline{w})\big]\Big). \tag{10}$$

Notably, the likelihood is undefined when any of the weights with a gamma prior are <0, ensuring that the weights which maximize the likelihood will always yield a monotonic model.

Care must be taken to use an optimization algorithm that appropriately handles undefined output from the objective function. Alternatively, in conjunction with the gamma prior, each w_j can be reparameterized to use a new set of variables $\overline{v}$ such that $w_j(v_j)=e^{v_j}$ for the weights with gamma priors and $w_j(v_j)=v_j$ for the rest. Now, instead of maximizing the likelihood by adjusting each w_j, the optimization algorithm adjusts each v_j. The mathematical details of this strategy yield the objective,

$$L=\sum_j\Big[(k-1)v_j-\frac{e^{v_j}}{\theta}\Big]+\sum_i\Big(t_i\log z_i\big(\overline{w}(\overline{v})\big)+(1-t_i)\log\big[1-z_i\big(\overline{w}(\overline{v})\big)\big]\Big), \tag{11}$$

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

where $\overline{v}$ is optimized by the gradient descent algorithm. Assuming that the gradient of $\overline{w}$ is computable, the gradient of $\overline{v}$ is computed using the chain rule,

$$\frac{\partial L}{\partial v_j} = \frac{\partial L}{\partial w_j} e^{v_j} + k - 1 + \frac{e^{v_j}}{\theta}, \tag{12}$$

for the the weights with a gamma prior and

$$\frac{\partial L}{\partial v_j} = \frac{\partial L}{\partial w_j}, \tag{13}$$

for the rest of the weights. Finally, the optimal $\overline{w}$ can be computed from the optimal $\overline{v}$ after training.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016