A Chemoinformatics Analysis of Hit Lists Obtained from High-Throughput Affinity-Selection Screening

NATHAN BROWN, HARTMUT ZEHENDER, KAMAL AZZAOUI, ANSGAR SCHUFFENHAUER, LORENZ M. MAYR, and EDGAR JACOBY

The high-throughput affinity-selection screening platform SpeedScreen was recently reported by the Novartis Institutes for BioMedical Research as a homogeneous, label-free screening technology with mass-spectrometry readout. SpeedScreen relies on the screening of compound mixtures with various target proteins and uses fast size-exclusion chromatography to separate target-bound from unbound substances. After disintegration of the target-binder complex, the binder molecules are identified by their molecular masses using liquid chromatography/mass spectrometry. The authors report an analysis of the molecular properties of hits obtained with SpeedScreen on 26 targets screened within the past few years at Novartis using this technology. Affinity-based SpeedScreen is a robust high-throughput screening technology that does not accumulate frequent hitters or potential covalent binders. The hits are representative of the most commonly identified scaffold classes observed for known drugs. Validated SpeedScreen hits tend to be enriched on more lipophilic and larger-molecular-weight compounds compared to the whole library. The potential for a reduced SpeedScreen screening set to be used in case only limited protein quantities are available is evaluated. Such a reduced compound set should also maximize the coverage of the high-performing regions of the chemical property and class spaces; chemoinformatics methods including genetic algorithms and divisive K-means clustering are used for this aim. (Journal of Biomolecular Screening 2006;123-130)

Key words: affinity-based high-throughput screening, SpeedScreen, chemoinformatics, chemoinformatics, scaffolds, genetic algorithms, subset selection, diversity

INTRODUCTION

The affinity-selection-based SpeedScreen methodology

SpeedScreen is a label-free, homogeneous, affinity-based high-throughput screening (HTS) technology with mass spectrometry (MS) readout, which was recently developed at Novartis.1,2 SpeedScreen has become one of the most frequently used HTS platforms at Novartis, especially when genomic targets with an unknown function (orphan targets) or targets with a known function but nontractable to a functional HTS have to be addressed. To achieve HTS performance despite the sequential MS readout, a primary SpeedScreen campaign is performed with pools of 400 compounds per well of the 96-well microplates. A SpeedScreen campaign starts with the incubation of the target protein with the pool of chemical compounds or natural products.

Protein-bound compounds are then separated from unbound compounds by fast size exclusion chromatography in 96-well format. Finally, the protein-ligand complex is disintegrated under the conditions used during a reversed-phase chromatography step, and the column effluent is directed to the ion trap MS detector with electrospray ionization.2 A software program enables fully automated comparison of the observed m/z values with the known molecular masses of the pooled compounds to identify binder molecules.

The size of a primary hit list from a SpeedScreen campaign with about 550,000 compounds depends on the target protein and ranges between several hundred and >10,000 compounds, which is in the same range as observed with functional HTS assays. In case of a large primary hit list (>2000 hits), the subsequent confirmation screen is performed with small pools of compounds (5-10 per well) to keep the analysis time within reasonable limits. In the final validation screen, single compounds are used throughout, which are analyzed in triplicate in the presence and in the absence of the target, respectively.

In the study reported in this article, both the primary and validated hit lists of 26 SpeedScreen campaigns were evaluated from...
signal-transduction proteins, enzymes, protein-protein interactions, proteases, kinases, and other nonmembrane proteins. The goal was to find and, eventually, attribute to sublibraries frequent SpeedScreen hitters and also to discriminate between sublibraries with a high hit rate and those without yielding any hits.

This analysis of the SpeedScreen hits is split into 2 distinct elements: a historical analysis of SpeedScreen campaigns, followed by an in silico investigation of the design of a reduced screening set. Since the inception of the SpeedScreen platform, a substantial number of assays have been performed, providing an enviable amount of quality data under consistent conditions. Therefore, by the application of this historical data in a retrospective data analysis, it is anticipated that instruction may be derived from the resulting information that allows us to further refine the screening protocols. Here, we report the data analysis of the historical data by use of a number of extant chemoinformatics data analyses together with novel informatics processes for the challenges that arise from this particular domain.

METHODS AND MATERIALS

SpeedScreen assays and analyzed HTS hit list

Each SpeedScreen assay results in up to 3 hit lists, with each following the other in a cascade based on triaging and validation, respectively. The 1st hit list resulting from an assay using pools of 400 compounds per sample is the primary hit list containing all molecules that are indicated to have bound to the protein. In some cases, the primary hit list is too vast to consider validating every one of the molecules using the validation protocol described above. Therefore, hit list triaging is performed in silico to triage this set of molecules by excluding an undefined number of molecules based on criteria such as scaffold diversity and bioavailability (absorption) indicators (see below); in some assays, hit list triaging was unnecessary. Because the number of hits returned from a primary hit list can vary considerably between assays, the enforcement of the triaging protocols is occasionally relaxed or enforced more rigorously on a case-by-case basis to provide a fairly constant subset of primary hits to pass onto validation. The result of the HTS triaging is referred to in this article as the triaged hit list. Last, the validated hit list contains those molecules that have been subject to the validation protocol described in the Introduction.

In total, this study considered a consolidated primary hit list of 34,054 molecules, a triaged hit list of 12,647 unique molecules, and a validated hit list of 1299 unique molecules over 26 SpeedScreen assays. Consolidated in this context refers to the reduction of all molecules in each of the 3 hit lists—primary, triaged, and validated—from each assay such that duplicates are removed but activity information is retained in the form of a binary vector with each position referring to whether the molecule binds (1) or does not bind (0) to the biological target used in that particular assay.

Chemoinformatics

The sheer quantity of data resulting from typical HTS campaigns necessitates the application of complex, and often pragmatic, data analysis methods to discover the information that is pertinent to allow the expert to derive indicators that can subsequently be applied in making informed decisions. The chemoinformatics techniques applied in the work reported in this article are described briefly below. The interested reader is referred to 2 recent introductory texts on chemoinformatics for more detailed information on the tools and techniques applied in this study and on chemoinformatics more generally and for a review of techniques directly applied in drug discovery.

Molecular descriptors. Given a molecular structure, a vast array of diverse properties can be calculated in silico very rapidly that express numerous characteristics of the molecule itself, generally called molecular descriptors. These descriptors can in turn be applied to perform various analyses on the molecules, thereby elucidating the salient aspects that are deemed important in the particular domain of interest under consideration. Although many thousands of descriptors have been reported, in this study, we use only a small subset of these descriptors. However, these descriptors are chosen for this analysis because they are applied widely, particularly using the protocols described below.

The rule of 5 defined by Lipinski and others is a set of 4 heuristics that have been defined such that they tend to indicate whether a molecule is likely to exhibit poor solubility or permeation. The heuristic rules are more than 5 H-bond donors, more than 10 H-bond acceptors, a molecular weight (MW) greater than 500, and a logP value of more than 5. The rule of 5 is so-called because each of the conditional criteria is a multiple of 5. Because the molecules in this article are those from HTS campaigns, it is arguable that rigid adherence to these heuristics is essentially a form of premature optimization, that is, omitting molecules too early in the discovery process that may transpire later to be of interest. Therefore, we apply a rule violation indication protocol that provides a fuzzier guide, thereby assisting in the filtering process. This mirrors the accepted practice of more than 2 rule violations being a basis for removal from the data set. We have also extended the rule of 5 to cover multiples of 7 of each of the properties defined by Lipinski and others (see above) as a less severe cutoff criterion, namely, 7 H-bond donors, more than 14 H-bond acceptors, greater than 700 MW, and a logP value of more than 7.

In addition to the standard molecular descriptors that are generally applied in chemoinformatics analyses, it is also possible to construct a novel descriptor vector for each molecule in which each position in the vector indicates whether that particular molecule was recorded as a hit against each respective target. This descriptor vector then provides an alternative representation and, using existing chemoinformatics techniques for dealing with structure-based descriptor vectors (or fingerprints), various analyses can be performed, such as similarity-based studies, the results of which are discussed later in this article.
Much of the research conducted in the field of chemoinformatics relies on the presumption that similar structures will tend to exhibit similar properties, in this case, biological responses. This is referred to as the similar property principle.\(^9\) Therefore, it is assumed that similar molecules will exhibit similarity values that are themselves similar when considering 2 different fingerprint representations. In conducting these analyses, we apply the biological profile vector together with the Extended Connectivity Fingerprint (ECFP) that is available in the PipelinePilot software published by SciTegic.\(^10\) With each of these binary vector representations, it is possible to calculate a similarity value between respective pairs of representations for 2 given molecules using a similarity coefficient, in this case, the Tanimoto coefficient. Therefore, it is possible to calculate the ECFP similarity and compare this to the biological profile vector similarity of the same 2 molecules. The hypothesis is that similar molecules (in ECFP space) will also exhibit similar biological activity profiles.

**Molecular scaffold frameworks.** The concept of a molecular scaffold is a chemically intuitive and often variable characterization of large collections of molecules. However, the application of a scaffold representation in a data analysis study such as this requires that a formal and consistent definition is used throughout the analysis process. In chemoinformatics, it is generally accepted that the definition of a molecular scaffold is achieved by application of the protocol defined by Bemis and Murcko\(^11\); the removal of side chains and the optional abstraction of all atoms and bonds to uncolored nodes and unweighted edges, respectively. These scaffold frameworks can then be applied in identifying common molecular shapes, allowing us to group or cluster these scaffolds and select representatives of these particular shapes in a way that is not necessarily possible with whole-molecule descriptor representations. This provides an alternative to the other in silico filtering techniques discussed above to ensure that the data sets are not overly enriched with representations of individual scaffolds.

**Library design and molecular diversity.** There are many published methods that are intended to assist in the design of molecular libraries for screening campaigns by suggesting a subset of molecules that should be considered for testing.\(^12\) Typically, it is accepted that a designed diverse subset of molecules is desirable because the intention of such a design is to cover the chemistry space represented in the entire data set. As with the determination of molecular scaffolds, however, there is currently no single rigorous definition of molecular diversity that is acceptable to all practitioners in the field of chemoinformatics—essentially a trade-off between covering the extremities of the space and representing the distribution of data points within that space.

A wide range of approaches have been published to select a diverse subset of molecules from a data set, for example, dissimilarity-based compound selection,\(^13\) cell-based methods,\(^18\) cluster analysis,\(^19\) onion designs,\(^20\) and neural networks.\(^21\) Here, we apply a clustering algorithm that permits the comparably rapid clustering of large numbers of data points marketed by Barnard Chemical Information\(^22\) using a hierarchical divisive K-means algorithm.\(^23\) From each cluster determined from the clustering algorithm used, 1 or more molecules are selected as being representative, thereby providing a representative subset of the chemistry space contained in the overall data set.

With the hierarchical divisive K-means algorithm, the entire data set is recursively divided into 2 by repeated application of the standard K-means algorithm, where \(K = 2\) in each case. In the application reported here, the entire cluster tree is generated (i.e., all clusters represent just a single data point), thereby permitting multiple levels of clustering to be induced from the tree. The hierarchical divisive K-means algorithm is of computational complexity \(O(N)\).

For this study, a genetic algorithm (GA) has been developed that selects a subset of plates by optimizing the coverage of the chemistry space represented by the entire library. Genetic algorithms have been applied to a wide range of scientific problems and provide a typically rapid method of discovering optimal or near-optimal solutions to problems that are otherwise insoluble in a realistic time frame (for a review of the application of GAs in the chemoinformatics domain, the reader is referred to the work by Clark\(^24\)).

The GA developed for the optimization strategy adopted in this study considers the selection of \(p\) unique plates from a larger pool of \(P\) plates such that the subset is optimized in terms of maximal coverage of the entire SpeedScreen library. The chromosome representation for subset selection is an integer string of length \(p\), where each gene in the chromosome may take any value in the range \(\{1 \ldots P\}\), ensuring there are no duplicate values. The crossover and mutation operators perform nonstandard functions when compared with the simple GA because the genes in the chromosomes are not order dependent.

In general, the crossover operator exchanges existing genetic material between \(2\) chromosomes. Therefore, in the GA for subset selection, it is sufficient to randomly select \(2\) genes from the respective parent chromosomes and exchange those between the chromosomes. This may be iterated over to exchange further genes between the chromosomes before the termination of a single crossover operation. The crossover operator requires a repair function to ensure that duplicate alleles are not present in the resulting chromosomes.

Mutation operators, however, operate by introducing new genetic material into the sampled search space. Therefore, it is reasonable that the allele of any gene on a given chromosome may be mutated to any valid allele value in the range \(\{1 \ldots P\}\) according to some probability. This operator also requires the application of the chromosome repair function to ensure there are no duplicate alleles in resultant operand.

The fitness function for the subset selection GA must calculate how well each candidate solution, or chromosome, covers the entire data set. To achieve this, we applied a clustering algorithm called divisive K-means that can rapidly, and with memory efficiency, cluster large data sets of long binary vectors, in this case,
FIG. 1. The number of unique molecules that are active against the specified number of targets for the primary, triaged, and validated hit lists, respectively. The sharp decay in the number of molecules illustrates that the given number of targets indicates that the SpeedScreen platform does not exhibit a severe frequent hitter problem.

FIG. 2. The number of unique molecular scaffold frameworks that are active against the specified number of targets for the primary, triaged, and validated hit lists, respectively. The decay of scaffolds that hit against the given number of targets falls away more slowly than with the molecules, but this plot still indicates that the SpeedScreen platform does not exhibit an issue in terms of promiscuous molecular scaffolds.

The approximately 550,000 data points contained in the SpeedScreen library as mentioned previously. The resulting exhaustive cluster tree may then be queried to return a specified number of clusters and their members from which each plate can be codified as a binary vector, with each point on the vector equating to either the absence (0) or presence (1) of that specific cluster. The more bits that are set in a given plate binary vector equate to a more representative plate in terms of the particular clustering analysis performed. Given that the optimization strategy must select p plates, each of the plate binary vectors may be combined using the bitwise logical OR operator to provide an overall indication of coverage for the given subset of plates. The role of the GA therefore is to evolve a subset of plates that maximizes this number.

RESULTS AND DISCUSSION

Analysis of frequent hitters

The frequent hitter analysis was performed on both the whole-molecule topological representation and the molecular scaffold framework to develop an understanding as to whether the SpeedScreen platform is susceptible to the frequent hitter problem. Figures 1 and 2 contain the results of whole-molecule and scaffold framework frequent hitter analyses, respectively.

In Figure 1, it is clear that the vast majority of the molecules (~70%-90% of the 3 hit lists) are present only in a single assay. This report soon tails off and leads to highly insignificant frequent hitters over the next 2 to 4 assays. Therefore, these results appear to demonstrate that, in terms of whole molecules, SpeedScreen does not suffer from a significant frequent hitter problem and can subsequently be used—based on this historical analysis—for a range of diverse screening campaigns.

A similar plot was generated for the frequent hitter analysis based on the molecular scaffold frameworks in Figure 2. Although the results show some very limited significance in the increase of frequent scaffolds over multiple assays, again, this is not deemed to be unduly concerning.

It can be observed that the scaffolds that hit against only 1 assay increase significantly from the primary hit list through to the validated hit list. The triaged hit lists collated from each of the SpeedScreen assays under consideration in this study have largely been subject to scaffold filtering, and this is evident in the 2-fold increase in relative single-hit scaffolds between the primary and triaged hit lists. However, this increase is followed by a further doubling of single-hit scaffolds between the triaged and validated hit lists, most likely as a result of applying a more rigorous screening protocol in the validation stage.

Analysis of molecular properties

The rules defined by Lipinski and others—the rule of 5 and their extension, the rule of 7—are applied to each of the primary, triaged, and validated data sets that resulted from all the SpeedScreen assays considered in this study and triaged to remove duplicates. The plots in Figure 3 indicate that although there are some limited numbers of rule violations, it is unlikely at this early stage in the drug discovery process that these molecules are unsuitable leads to be considered downstream; it is important, however, that the leadlikeness of molecules should also be observed as a selection criterion. What is perhaps more interesting is the apparent signal that the validated set of molecules tends, in both the rules of 5 (Fig. 3) and 7 (Fig. 3, inset), to result in more rule violations than for the primary and triaged sets. Although, again, it is important to place the application of these rules in the HTS context, it does go...
some way to validating our hypothesis that the larger and more complex molecules are more likely to bind, and this is concomitant with the mechanisms of binding. Because of the MS detection, which will completely miss ligands that have covalently bound to the target, the SpeedScreen platform is not adversely affected by covalent molecules, as can be the case with ALARM NMR.30

It is not believed that when applying these heuristic rules as indicators, the resulting number of rule violations should generate too much concern when passing these molecules on to chemists. Indeed, it can be suggested that these larger, more complex molecules should be studied more closely to ascertain whether there are any particularly interesting structural moieties that could be helpful in structural rationalization through further in silico modeling, possibly applying both predictive modeling and iterative molecular fragmentation or pruning methods. This can be seen as an inverted version of the SAR by NMR methodology,31 in which the aim is to build novel ligands from building blocks that exhibit desired binding affinities.

**HTS plate analysis**

Each of the plates used in a current SpeedScreen assay has been designed by a complex stacking based on pooling compounds from the existing library plates to achieve the 400 compounds per well that are used for primary SpeedScreen assays. Because these plates have been stacked in this method, it is expected that there will not be any significant artifacts of design in terms of the number of hits varying greatly between each of the plates. Figure 4 illustrates the percentage of primary and validated hits that occur on the plates together with the frequency of plates that exhibit this proportion of activity.

Using the descriptors reported by Lipinski and others8 for the rule of 5 heuristics, it was found that there was a pronounced difference only in property distributions in the MW and logP descriptors—reported in Figures 5A and 5B, respectively—suggesting that the preferred molecules are larger and more lipophilic, which concurs with the concepts of molecular binding.

This analysis demonstrates that although there is some enrichment of actives on some of the plates, it is very insignificant, thereby suggesting that a designed subset would be best achieved simply by attempting to maximize the coverage of the entire SpeedScreen library based on cluster analysis.

**Correlations of biological profiles with structure**

As has been noted previously, it is a reliable heuristic that similar structures will exhibit similar properties, in this case, biological activities.32 To this end, random pairs of molecules were sampled from the consolidated primary hit list, and the Tanimoto similarity was calculated between their biological activity profiles and structure fingerprints, respectively (repeated 5000 times); the resulting correlation plot is given in Figure 6. In terms of the SpeedScreen platform, this is a promising analysis because it suggests that this method is detecting molecules that otherwise would have been missed through the application of similarity searching alone. It should be noted that the striped appearance of the biological similarities is an artifact due to the limited number of dimensions in the biological activity space.

**Optimized selection of a subset of plates using a genetic algorithm**

To test the efficacy of the GA in terms of its ability to evolve a subset of plates that in some measurable way maximizes the coverage of the entire library, a number of experiments were conducted.
with various numbers of clusters extracted from the divisive K-
means package: 4096, 8192, 16,384, and 32,768 clusters, respec-
tively. In each case, the aim is to maximize the number of each of
the sets of clusters that are represented in the final subset design.
Because GAs are stochastic optimization strategies, each particu-
lar experiment was conducted 5 times to determine the degree of
repeatability of the optimization.

However, what is not clear, particularly in the domain of
SpeedScreen campaigns, is whether a designed subset will per-
form significantly better over a random selection strategy. There-
fore, we provide a comparison of our designed results together
with 5 sets of 30 random subsets each to validate our method.

A representative sample of the evolution rate is provided in Fig-
ure 7, reporting each generation as the minimum, mean, and max-
imum fitness over the 5 individual GA runs for the optimization
from 8192 clusters. These figures illustrate atypical fitness ascent
for GAs, in which the norm tends to be a rapid increase to a
suboptimal solution in very few generations and a gradual increase
in fitness thereafter. Indeed, it can be observed that there is little—
in general, less than 5%—improvement in coverage over the start-
ing conditions throughout the experiments.

A comparison of the designed versus random subsets is given in
Table 1 in terms of coverage of the respective clusters, and it dem-
onstrates that there is little enrichment in coverage between the 2
methods. However, this is not to say that a designed subset in this
case may not be beneficial in practice because the method is indi-
cating increased coverage. What is not clear is whether this minor
increase in coverage is substantial enough, given the method by
which we determine coverage, to warrant applying this relatively intensive design process in practice. The GA developed in this study, although relatively unimportant for the optimization problem at hand, has nevertheless demonstrated that given the sets of plates, it is unlikely that applying a complex and computationally intensive subset design method rather than the much more rapid random subset selection method will be of benefit. However, this does not negate the applicability of the subset selection GA reported in this study to additional subset selection problems.

CONCLUSIONS

This study has investigated the potential for artifacts to occur in the SpeedScreen methodology as a result of screening large numbers of compounds per well. However, as has been demonstrated in this study, the SpeedScreen platform appears to be largely devoid of these types of limitations. This is due to both the in silico filtering methods applied to filter and triage the resulting primary hit lists and also to the rigorous screening process at the validation stage. One expectation was that common scaffolds would occur frequently in the validated structures, but this has been demonstrated not to be a significant problem in SpeedScreen, most likely as a result of molecular- and scaffold-based triaging of the primary hit lists.

A substantial proportion (42%) of the molecules from the validated hit lists satisfy all of the Lipinski rule of 5 heuristics, whereas most satisfy at least 2 of the rules regarded to be sufficient adherence to permit continued consideration downstream. The most significant heuristics that are violated are those concerning MW and logP; this is in line with the understanding of nature of the under significant heuristics that are violated are those concerning MW and dated hit lists satisfy all of the Lipinski rule of 5 heuristics, whereas

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Random</th>
<th>Designed</th>
<th>Absolute Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>4096</td>
<td>3960.40</td>
<td>6.19</td>
<td>4008.80</td>
</tr>
<tr>
<td>8192</td>
<td>7603.40</td>
<td>17.59</td>
<td>7763.60</td>
</tr>
<tr>
<td>16386</td>
<td>14215.40</td>
<td>101.99</td>
<td>14747.40</td>
</tr>
<tr>
<td>32768</td>
<td>26027.40</td>
<td>126.88</td>
<td>27075.00</td>
</tr>
</tbody>
</table>

Overall, the relative similarity of both random and designed diversity subsets suggests that the SpeedScreen library contains sufficient representation in terms of diversity that a designed diversity subset offers little over random.

ACKNOWLEDGMENTS

Dr. P. Acklin, Dr. R. Amstutz, Dr. C. Engeloch, Dr. P. Fuerst, N. Lehmann, Dr. U. Schopfer (all Novartis Institutes for BioMedical Research), and Professor P. Willett (University of Sheffield) are acknowledged for various discussions and support. N.B. is funded by a Presidential Postdoctoral Fellowship from the Novartis Institutes for BioMedical Research.

REFERENCES