Development and Analysis of a Plate Based Diversity Set

Jens Loesel

Pfizer Global Research & Development, Sandwich Laboratories, UK

ABSTRACT

A full file HTS is the only way to cover all of a drug companies compound space in its entirety. For many purposes a full file HTS seems neither cost efficient nor necessary. The work described here investigates how much of the Pfizer compound space can be covered using a sub selection of existing plates. Based on this work a prioritised plate order has been proposed.

INTRODUCTION

The aim of this work was an investigation of the Pfizer compound space covered by the current full file HTS. The final goal was a proposal of an ordered plate list to maximize the coverage of chemical space with a subset of the file.
The aim of the work is quite similar to work carried out at Novartis described in "Plate Cherry picking": A Novel Semi-Sequential Screening Paradigm for Cheaper, Faster, information-Rich Compound Selection [1]– Crisman at al [1].
The poster describes the work done to derive the subset and compares differences in methodology to the Novartis work.

Definition of chemical space

There are multiple ways how to describe the chemical space. Discussion of the merits of each approach would go beyond this poster.
For pragmatic purpose we chose a 6 dimensional BCUT space calculated using the DVS software with 8 bins along each axis.

Main reasons for the choice of chemical space were
- Readily available
- Prior work done using the same space for a cherry picked compound selection [2]
- Preference of a bin-based definition of chemical space

Exclusion of cells with very low coverage from optimization

Diverse selections in chemical space without special safe guards or filters often lead to an enrichment in non-drug like compounds. We had to ensure that the goal of a maximum coverage would be reached without compromising the quality of the compounds selected.
To check the quality of outliers we did a visual inspection of 25 diverse compounds from cells with occupancy 1.
>80% of the selected compounds were regarded as non drug like. We are aware that the diverse selection did amplify the problem. Based on these findings we felt justified to exclude all cells with coverage <10 from the optimization process.

Non random distribution of combinatorial compounds

Another issue we had to take into account was the non random distribution of combinatorial compounds.
Figure 1 shows that approx. half of all plates in the screening collection are dominated by combinatorial derived compounds while the other half is mostly devoid of such compounds.

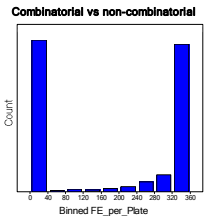


Figure 1: Plate distribution of combinatorial derived compounds

Quality of individual plates

To further enhance the quality of the selection and to guide the selection process we checked the overall quality of each individual plate.
A similar approach to Lipinski's work on the Ro5 was used for the property distributions of the plates. Cut-offs for plate properties were derived which included 90% of the 'best' plates. We will refer to cut-offs derived this way as the Rule of 40.

a. More than 200 unique compounds on a plate
b. More than 160 compounds with zero Ro5 violations
c. More than 120 compounds free of less desirable substructures

Refined goals for the plate selection

Based on the definition of chemical space using BCUT descriptors in a six dimensional space with the exclusion of low coverage cells and we set our final goals to

- Maximum overall coverage of number of cells
- 95% double coverage of each bin to allow coverage even in case of experimental error
- half the plates should be selected from predominately combinatorial plates and the other half from non-combinatorial plates
- plates with druglike compounds preferred, no exclusions

Optimization Workflow

1. Randomization of all plates not yet selected with penalty for non Ro40 plates
2. Split in combinatorial and non combinatorial plates
3. Order all plates according to coverage with combinatorial and non combinatorial plates alternating
4. In each loop keep 100 additional plates until enough plates selected
5. Check for convergence
6. Final plate order

Repeat if needed

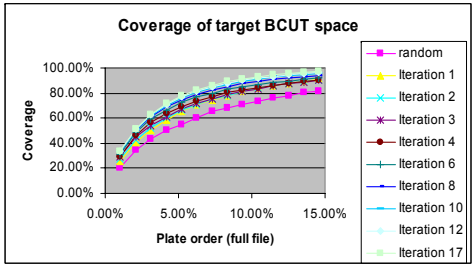


Figure 2: Convergence of plate coverage

Coverage of low occupancy cells

The table below investigates the coverage of low occupancy cells. Low overall (<50%) coverage is colored in orange.
Random numbers where derived calculating statistical expectation values based on the number of cells with occupancy n and the number of plates selected.
It should be noted that the selection process did result in a positive bias in coverage of these cells over random despite exclusion of them in the optimization process.

Cell occupancy	Percentage cover	Increase on random	Percentage double cover	Increase on random
1	18.67%	58.3%	NA	NA
2	34.44%	58.7%	3.72%	49.1%
3	46.29%	53.2%	10.60%	64.6%
4	55.42%	47.1%	18.91%	68.3%
5	63.41%	43.1%	27.28%	64.7%
6	69.87%	40.5%	34.35%	58.3%
7	73.52%	34.7%	42.93%	60.2%
8	79.14%	32.6%	43.88%	34.2%
9	82.30%	30.3%	52.70%	41.9%

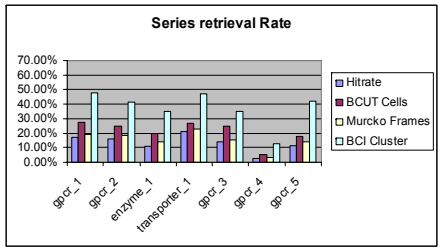
Table 1: Coverage of cells with low occupancy

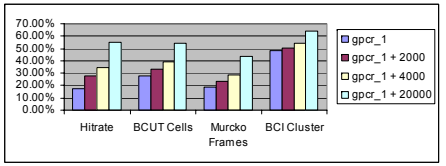
Comparison with Novartis workflow

We opted for a different description of chemical space. This resulted in a smaller number of cells compared to Murcko Series. In addition we excluded low occupancy cells to prevent enrichment in non druglike compounds.
Possible Coverage of space depends mainly on the definition of the space as well as the number of plates chosen. The smaller number of cells compared to the Novartis workflow results in a higher relative coverage. We have not recalculated coverage using Murcko Series but would expect similar results.
We used a GA like approach in our optimisation. We felt this was necessary to avoid any local minima resulting from the primary order of plates. Due to the smaller number of cells to cover there is an increased chance to cover the same space with multiple different cells.
Due to the diverse nature in the quality of plates we derived the rule of 40. The plate selection is heavily order dependent. The penalty for non Ro40 effectively excludes this plates – unless they offer some unique coverage of plates not on other plates.
We treated plates with predominantly combinatorial plates in a preferred way by ensuring 'equal selection' for them. We could have reached a higher coverage without the preferred handling of these plates.

Recall for active series

We also analyzed the results using Murcko Frames. The graph below shows the recall of scaffolds using Murcko Frames using approx. 12% of the total file.
The graphs show a very even rate of recall. Using a Bayesian Model similar to the Novartis approach (second graph) confirm yields comparable enrichment in retrieval rate.





Discussion

We have shown here an alternative approach to derive a Plate Based Diversity Set. We agree with their work that 'Plate Cherry Picking' can be a faster and more cost effective way of screening compared to a full file HTS if followed up by triage and computational models.
We also show that an optimized selection of plates is superior to a random selection – both in coverage of chemical space as well as recall of series.
The non random distribution of compounds on plates due to historical workflows need to be handled in an individual manner. We insisted on an equal distribution of combinatorial and non combinatorial compounds. In addition we derived the Ro40 as a mean to prioritize certain plates.
Different bias due to other history will need alternate solutions.

ACKNOWLEDGMENTS and REFERENCES

Thanks to David McLoughlin and his colleagues in the HTS CoE for the discussions and the help in performing this work. Thanks also to Kuen Yeap for BCUT calculations and Alexander Alex for various input.

References:
[1] "Plate Cherry Picking": A Novel Semi-Sequential Screening Paradigm for Cheaper, Faster, Information-Rich Compound Selection, Crisman,T.J.;Jenkins,J.L.;Parker,C,N.;Hill,W.H.L.;Bender,A.; Deng,Z.; Nettles,J.H.;Davies,J.H.;Glick,M. *J.Biomol Screen 2007; 12; 320*
[2] Designing compound subsets: comparison of random and rational approaches using statistical simulation, Yeap, S.K.;Walley, R.J.;Snarey, M.;van Hoorn, W.; Mason, J.S. *JCIM* submitted

Chemoinformatic Conference
Sheffield 18-20 June 2007