CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **86** (4) (2013) 435–441.
http://dx.doi.org/10.5562/cca2301

Original Scientific Article

A Heuristic Algorithm for Plate SelectionThat Maximizes Compound Diversity†

Hongyao Zhu,* Jacquelyn Klug-McLeod, and Gregory A. Bakken

Pfizer Worldwide Research and Development, Groton, Connecticut, United States

RECEIVED JUNE 7, 2013; REVISED SEPTEMBER 12, 2013; ACCEPTED SEPTEMBER 27, 2013

Abstract. A heuristic algorithm was developed to maximize compound diversity within a subset of screening plates used in high throughput screening (HTS) to initiate the drug discovery process. The approach overcame the challenge of combinatorial explosion for selecting plate subsets with maximum compound diversity. The method yielded novel forms of plate-based diversity subsets (PBDS) for HTS screening in lead discovery. The algorithm, its validation and the application to our screening collection are outlined. (doi: 10.5562/cca2301)

Keywords: high throughput screening (HTS), plate-based diverse subset (PBDS), combinatorial optimization.

INTRODUCTION

As an effective approach to lead discovery, high throughput screening (HTS) is widely used in industry and academia to initiate the drug discovery process.[1–6] However, for large compound collections, HTS is time-consuming and raises cost-effectiveness concerns. In order to address this challenge, it is natural to screen subsets of large screening collections. A subset can be selected on the basis of target-specific,[7] gene family-specific[8] or chemical diversity for wider coverage of chemical series.[9,10] Beginning in 2006, a novel form of subset screening coined plate-based diversity screening (PBDS) was developed and implemented in Pfizer.[11] Novartis reported on similar implementations in 2007[12] and 2009.[13] The PBDS offers a unique approach in that the screening subset was constructed by plate selection as opposed to cherry-picking individual compounds.

In 2009, a second generation set, PBDS2,[14] was constructed and updated with an improved coverage of the chemical space for Pfizer's new and enlarged screening file, whilst only selecting the same number of plates for screening. This became one of the default singleton (one compound per well) HTS subsets for lead discovery and has been routinely used to validate assays in Pfizer. By design, the first 100 plates of PBDS2 were the most diverse. As a result, for various reasons, these plates were more often favored and screened over the entire set. Due to this uneven use of the PBDS2 set, the supply of the first 100 plates was being quickly depleted.

In order to further improve the cost-effectiveness and the efficiency of our PDBS-based HTS approach, a different strategy for the PBDS subset selection was essential. This led to development of PDBSx. The goals for the selection processes are:

(1) Create a series of PBDSx subsets by partitioning the singleton plates such that each of the sets is designed to be equivalent with each other set in terms of overall compound diversity. This allows for rotation of subsets to screen and, therefore, ensures more even depletion of the file in longer term use.

(2) Ensure each of the sets is rank ordered in terms of overall compound diversity.

To fulfill these requirements and considerations, we defined the metric that can be used to reflect the diversity of plates, established a novel algorithm of combinatorial optimization for selecting or partitioning the collections of plates, and conducted validation of the approach and used it in the final selection of plate subsets. The detailed method is described in section 2 and the results are discussed in section 3.

† Dedicated to Professor Douglas Jay Klein on the occasion of his 70th birthday.

* Author to whom correspondence should be addressed. (E-mail: hongyao.zhu@pfizer.com)

METHODS

Significant changes and improvements to the Pfizer HTS screening file were made in order to reduce the file size by removing duplicate batches and eliminating replicate samples, redundant compounds and unattractive compounds[15] in considering drug/lead like molecular properties.[16–18] Plates were produced by the Liquid Store in Pfizer. Each 16 by 24 plate contained 360 wells for test compounds, with 24 wells reserved for controls. There were ~2.87 million compounds in 7977 plates. Accelrys Pipeline Pilot's ECFP4 fingerprints[19,20] were used to compute compound pairwise similarity. Though compound pairwise based similarity has been effectively used in a wide variety of scenarios[21,22] and corresponding uses are straight forward, it was not feasible to directly work on all of more than 4 trillion compound pairs. Additionally general compound clustering methods could not be directly used for such a huge dataset. We were not aware of any internal or published algorithms available to be adopted directly for plate selection to meet our needs**.** In this section, we introduce a stepwise heuristic algorithm to select equivalent subsets of plates.

Metric for Diversity of Plates

There are different approaches to describing the chemical space coverage for a screening file. Novartis first used Murcko frames and chemical fingerprints to quantify the coverage of space in their reports of a plate-base diversity selection.[12,13] Pfizer used a cell-based method to quantify the coverage in lower dimension BCUTs[23,24] chemistry space for diverse subset[10] and earlier generations of PBDS subsets.[11,14] In this report, the ECFP4 fingerprints-based similarity is used to determine the overall diversity of the compound collection.

On the plate level, the metrics for diversity are defined as intra-plate similarity (P_{ii}) for all compound pairs within a plate (64240 pairs) and inter-plate similarity (P_{ij}) for all compound pairs between two plates (129600 pairs) by counting compound pairs with Tanimoto similarity greater than 0.70. A higher cut-off value of Tanimoto similarity results in a lower resolution to plate diversity measurement and yields a very sparse matrix of pairwise similarity, while a lower cut-off introduces false positive to plate diversity measurement. Based on the study of 0.05 similarity interval scanned on [0.60, 0.90] for a subset of 6 plates, we chose 0.70 as the optimal Tanimoto cut-off in this work. The ECFP4 fingerprints were calculated with Pipeline Pilot 8.5.[20]

Methods for Plate Selections

Selecting a subset of plates that minimizes the overall compound similarity ensures a wider coverage in terms of compound diversity. For our collection of plates, using an exhaustive search is not feasible due to the combinatorial explosion. For example, selecting 1300 out of 8000 plates would require searching through more than 10^{500} combinations from the collection of plates. Therefore, we introduced a stepwise method below.

The symbols used in this report are described below:

- (1, 2, 3, …): plate index;
- (*A, B, C, ...*): PBDSx and alternative subsets;
- *P*: set of entire collection of plates;
- S_k^{α} $S_k{}^{\alpha}$: set of selected subset with size *k*;
- V_M: set of available plates (plate pool) with size *M*.

A PBDSx subset is selected as:

$$\min(\sum_{i \in S} P_{ii} + \sum_{i \neq j;\ (i,j) \in S} P_{ij}) \quad (1)$$

where *S* is the selected subset. One plate is selected at each step for a corresponding plate subset following the rules:

- Intra-plate diversity is given a higher priority for selection:
  (1) when a new plate is added to S_k, the plate itself is required to be highly diverse to ensure overall diversity for S_{k+1};
  (2) in each step, a subset of available plates with the most diversity V_M is created as a pool of plates for selection.
- Inter-plates diversity is optimized by using:

$$\min(\sum_{j \neq i} P_{ij}),\ i \in S_k \quad (2)$$

- In each step, alternative sets are selected so that diversity measures within each of the sets are as close as possible. Therefore all alternative sets would be equivalent with each other in terms of overall compound diversity.
- After a plate is selected, it is removed from further consideration in the selection process. Then a new V_M is created from the available plates for the next iteration.

Thus the process for plate selections can be described as formulas (3) which are for the first plate in each PBDS subset and formulas (4) for following plates to be selected to add to the subset.

$$A_1 = \min(\sum_{j \neq i} P_{ij}),\ i \in V_M; j \in P$$
$$B_1 = \max(\sum_{j \neq i} P_{ij}),\ i \in V_M;\ \ j \in S_k^A \quad (3)$$
$$C_1 = \max(\sum_{j \neq i} P_{ij}),\ i \in V_M;\ \ j \in S_k^A, j \in S_k^B$$

$$A_n = \min(\sum_{j \neq i} P_{ij}),\ i \in V_M; j \in S_k^A$$
$$B_n = \left[\max(\sum_{j \neq i} P_{iA}), \min(\sum_{j \neq i} P_{ij}) \right],\ i \in V_M; j \in S_k^B \quad (4)$$

where n > 1.

For the first plate of the first set, A_1, an available pool of plates, V_M is defined based on intra-plate diversity. A single plate is then selected based on minimization of that plate's similarity relative to remaining plates. The first plate of the second set, B_1, is selected from the pool V_M of plates with high intra-plate diversity, while also ensuring maximum similarity to A_1. The first plate of the third set, C_1, is selected by considering intra-plate diversity as used for A_1 and B_1 selections, while ensuring a plate with maximum similarity to the already selected sets A_1 and B_1.

In the subsequent steps to add a plate to set A, A_n is selected from the available pool of plates (V_M) with maximum overall inter-plate diversity within the selected subset. For subsequent plates to set B, B_n is selected from V_M maximizing similarity to A_n and also maximizing overall inter-plate diversity for B. The selection procedure for other alternative subsets (C, D, etc.) is the same as that of subset B.

This plate selection algorithm is illustrated schematically in Figure 1, in which each dot represents a plate, and the distance between dots mimics represents the similarity measures for pairs of plates. In each step of selections, a plate is added to a corresponding subset (A, B, or C) so that the plate is far away or dissimilar to the selected subset but close or similar to corresponding plates in other subsets. This, therefore, yields equivalency for the alternative PBDS subsets in terms of compound diversity within each respective subset.

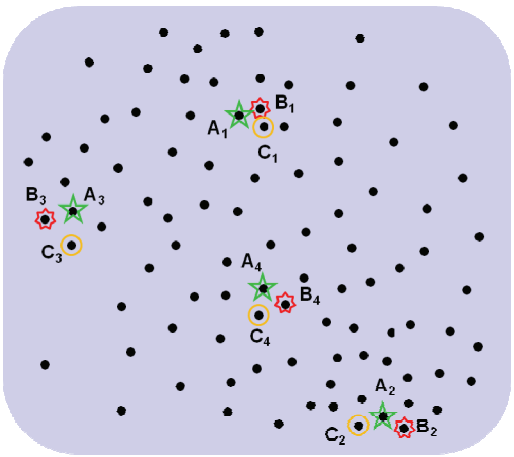

Figure 1. The plates which are chosen at each of the selection steps for perspective subset. Each dot represents a plate. The position and distance illustrate the similarity between plate pairs.

Implementation for the Methods of Plate Selections

The plate selection algorithm was implemented in Python 2.4.3 running in Red Hat Linux version-2.16.0. A pilot study with 5417 plates and the actual PBDSx application with 7977 singleton plates were used for algorithm validations, which demonstrated the algorithm outlined in 2.2 has a computational efficiency of ~ $O(N^2)$.

RESULTS

The list of 7977 screening plates available for experimental screening was obtained from the materials management group in Pfizer as of October 2011. Each 16 × 24 plate contained up to 360 wells with 40 µL and 30 mM of compounds and 24 control positions. Information for each individual plate contained compound identifiers mapped to well positions. The corresponding structure information was retrieved from the Pfizer internal compound database as SMILES strings.[25,26]

As described in section *Metric for diversity of plates*, the metrics of plate diversity were measured by counts of similar compound pairs for all intra- and inter-plate compound pairs. The distribution of plates in terms of intra-plate similarity is shown in Figure 2. It should be noted that about 25 % of the 7977 screening plates were constructed with high intra-plate diversity with the number of similar pairs less than 20, while the rest of the plates were distributed with widely ranging intra-plate diversity. Because of such distribution characteristics, we identified an optimal size for the set of available plates (plate pool) V_M in our optimization process to comprise the maximal diversity and acceptable computational demands for selecting PBDSx subsets.

Instead of selecting a single PBDS subset as previously implemented, we decided for PBDSx to partition 7977 plates into 6 subsets with equivalent diversity in each subset by applying the algorithm as described in *Methods*. Based on the usage experience of previous versions of PBDS and our screening capacity, the size of 1250 ~ 1500 plates was desirable. Therefore, 6 subsets were chosen for partitioning to yield PBDSx. In each subset, the plates were rank ordered in terms of the total diversity of the subset with the first plate of most

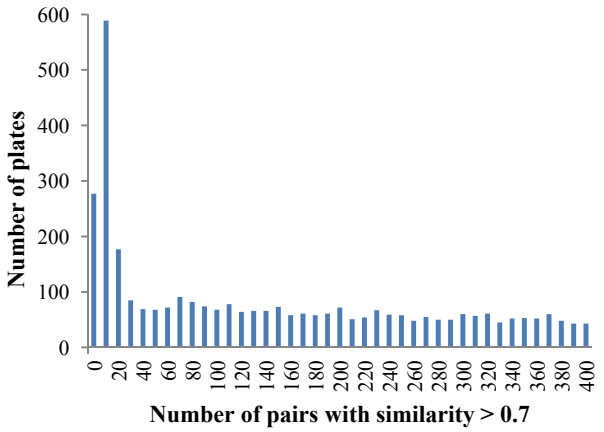

Figure 2. Histogram of the number of screening plates in terms of intra-plate similarity measured by counts of similar compound pairs within each plate.

diversity and following plates with decreasing diversity. The overall diversity plots for 6 subsets at each selection step are shown in Figure 3 with different sizes of selection pool (V_M): 150, 450, 2000, and 4000. It is demonstrated that with increasing selection pool size, all 6 subsets had a strong tendency to be more equivalent in diversity. For pool sizes of 2000 and 4000, the diversity variations between 6 subsets became negligible. When the pool size reached 2000, ~ 25 % of the plates with higher intra-plate diversity (*i.e.*, there are less than 20 pairs having similarity greater than 0.7 as indicated in Figure 2,) were included in the selection pool to ensure an optimal plate selection for compound diversity. A large size of the selection pool M was required to give sufficient selection options to maximize the overall diversity at each step in the selection process. Based on the results shown in Figure 3, selection pool size of M = 4000 was used for V_M in our final optimization for the selection of PBDSx subsets.

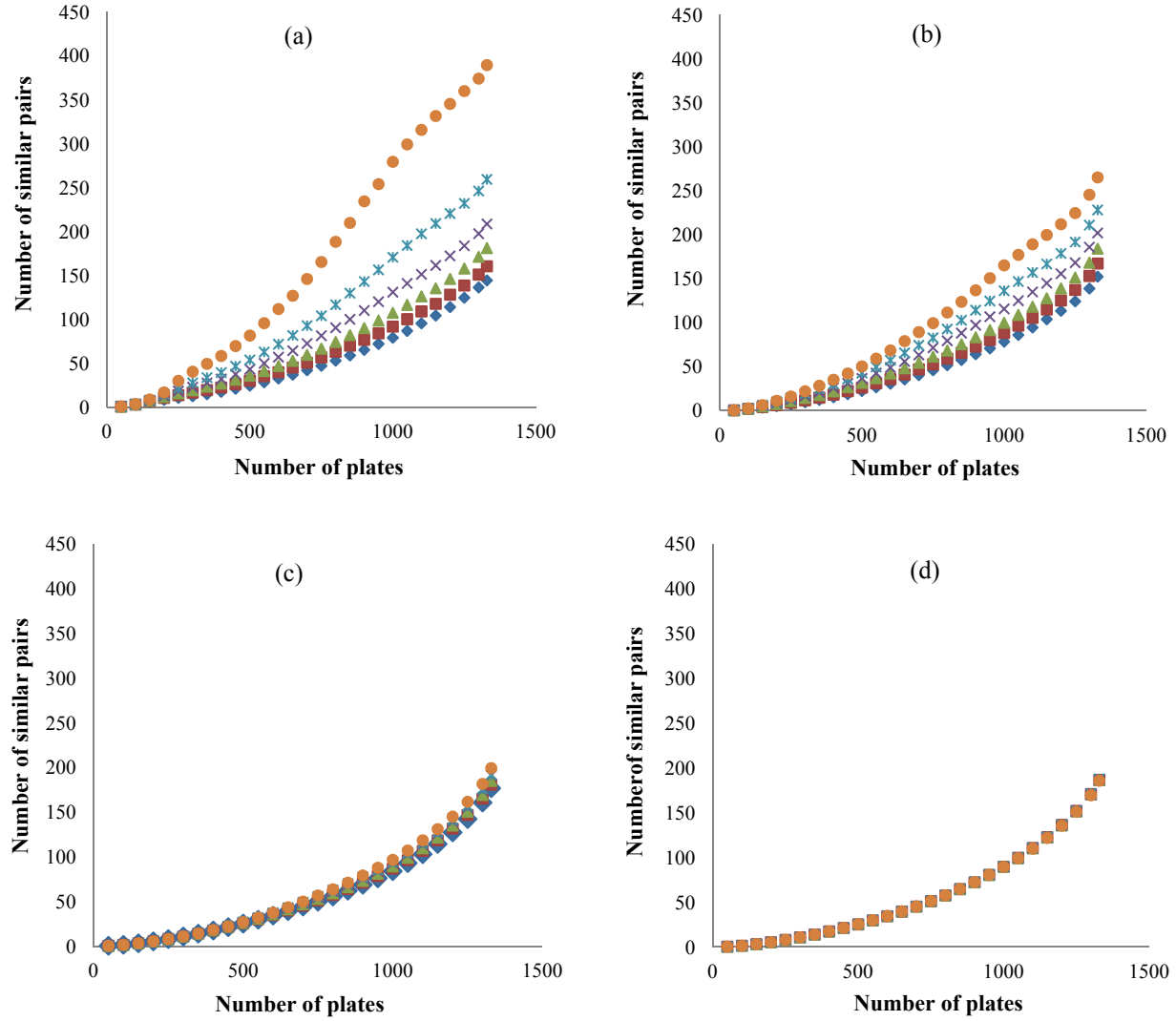

Figure 3. The diversity distributions of six PBDSx subsets depend on the size of selection pool of available plates. The pool size M are: (a) 150; (b) 450; (c) 2000; (d) 4000, and y-axis in each plot is in thousands unit.

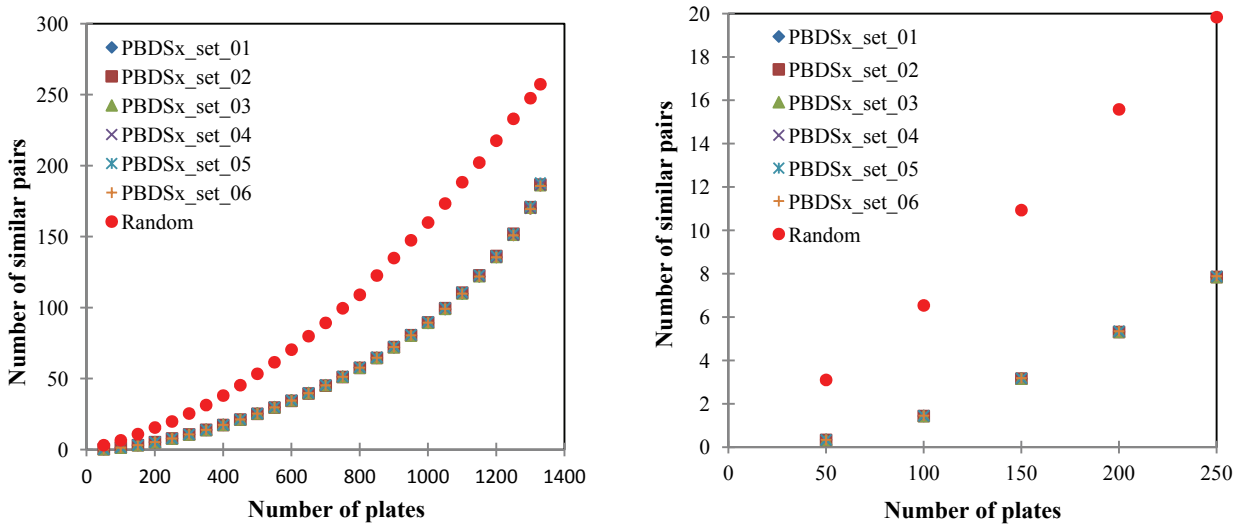

Figure 4. The diversity distributions of six PBDS subsets with the pool size 4000. The *y*-axis in each plot is in thousands unit.

Table 1. Total number of compound pairs with Tanimoto similarity greater than 0.70 for the selected six PBDS subsets

Number of plates	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
50	341	335	335	332	343	340
100	1442	1441	1439	1440	1456	1456
150	3168	3163	3166	3167	3186	3187
200	5320	5316	5321	5327	5353	5351
250	7849	7851	7848	7866	7884	7892
300	10727	10722	10725	10731	10767	10767
350	13908	13910	13912	13919	13974	13955
400	17424	17358	17435	17402	17509	17458
450	21244	21134	21301	21184	21337	21268
500	25386	25208	25457	25295	25489	25379
550	29861	29639	29917	29757	29993	29822
600	34652	34398	34707	34570	34819	34615
650	39799	39538	39862	39785	40010	39791
700	45311	45089	45401	45376	45580	45325
750	51292	51131	51352	51371	51581	51312
800	57715	57597	57729	57835	58055	57783
850	64615	64609	64651	64842	65088	64766
900	72092	72205	72188	72471	72692	72309
950	80288	80537	80387	80735	80958	80549
1000	89236	89622	89379	89646	90025	89461
1050	99000	99578	99174	99396	99925	99165
1100	109887	110579	110022	110158	110869	109856
1150	121936	122691	122166	122141	122984	121882
1200	135634	136362	135672	135492	136492	135496
1250	151182	151998	151382	151096	152173	150950
1300	169943	170723	170457	169964	171403	169397
1329	186263	186609	186357	186313	187880	185734

In Figure 4, the diversity distributions of six PBDS subsets are plotted for each of the plate selection steps with pool size M = 4000. For comparisons, 20 randomly selected subsets were generated from the collection of 7977 plates. The mean diversity distribution of the 20 random subsets is also shown in Figure 4a. The same plots are zoomed-in as shown in Figure 4b to reflect the details for early steps in plate selection. Comparing to the random selections, our algorithm provided optimal results in terms of compound diversity in selected subsets. For example, the mean value of total number of compound pairs with Tanimoto similarity greater than 0.70 for the 6 subsets were 338 for M = 50, 25, 369 for M = 500, and 89, 562 for M = 1000, compared to the mean of random selection 3104 for M = 50, 53, 410 for M = 500, and 159, 994 for M = 1000, respectively. This might imply that the stepwise approach significantly lowered the possibility of trapping in false minima in the selection process. The total numbers of compound pairs with Tanimoto similarity greater than 0.70 for the six selected subsets are listed in Table 1. As illustrated in Figure 4 and Table 1, the selection algorithm outlined in this report has been validated in comparing to the negative control of random selection of plates.

CONCLUSION

To overcome challenges of combinatorial explosion for diverse subsets in plate selection, a heuristic algorithm for plate selection that maximizes the diversity of compounds in selected subsets of plates was developed and validated. The method was utilized to partition Pfizer's 7977 singleton plates into 6 equivalent PBDSx subsets in terms of overall compound diversity within a subset. There were 1330 plates in partitions 1, 2 and 3, and 1329 plates in partitions 4, 5, and 6.

The stepwise plate selection approach maximizes the overall intra- and inter-plate diversity. Since the 6 PBDSx subsets have equivalent diversities, they can be used on a rotating basis for screening. This strategy ensures more even depletion of the screening file over a longer term course. Within each subset of PBDSx, the plates are ranked by diversity, and the top-N plates together are the most diverse partial subset in counting of both intra- and inter-plate diversity while the bottom plates give less diversity contributions. Thus, the partial subset can be directly used to serve different purposes in HTS screening. Another consideration could be a combination of partial subsets; for example, taking the top 200 plates from each of the 6 subsets yields to generate a set of 1200 plates for HTS screening. In this case, we must be aware of the inter-set similarity that would occur due to the selection process.

The method outlined in this report yielded a novel, third-generation PBDSx which has been used as one of the standard sources of singleton-HTS-based lead discovery in Pfizer since 2011.

Acknowledgements. We gratefully acknowledge the following Pfizer colleagues in creating the screening file: Rosalia Gonzales, Travis Mathewson, Donna Tosta, Holly McKeith, Keith Miller, and Peter Tunucci.

Abbreviations. ECFP4, SciTegic/Accelrys' level 4 extended connectivity fingerprints; HTS, high throughput screening; PBDS, plate-based diversity subset.

REFERENCES

1. B. Cox, J. C. Denyer, A. Binnie, M. C. Donnelly, B. Evans, D. V. Green, J. A. Lewis, T. H. Mander, A. T. Merritt, M. J. Valler, and S. P. Watson, *Prog. Med. Chem.* **37** (2000) 83−133.
2. J. W. Davies, M. Glick, and J. L. Jenkins, *Curr. Opin. Chem. Biol.* **10** (2006) 343−351.
3. R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, and G. S. Sittampalam, *Nat. Rev. Drug Disc.* **10** (2011) 188−195.
4. L. M. Mayr and D. Bojanic, *Curr. Opin. Pharmacol.* **9** (2009) 580−588.
5. M. C. Monteiro, M. de la Cruz, J. Cantizani, C. Moreno, J. R. Tormo, E. Mellado, J. R. De Lucas, F. Asensio, V. Valiante, A. A. Brakhage, J. P. Latge, O. Genilloud, and F. Vicente, *J. Biomol. Screen.* **17** (2012) 542−549.
6. D. A. Pereira and J. A. Williams, *Br. J. Pharmacol.* **152** (2007) 53−61.
7. R. H. Shoemaker, D. A. Scudiero, G. Melillo, M. J. Currens, A. P. Monks, A. A. Rabow, D. G. Covell, and E. A. Sausville, *Curr. Top. Med. Chem.* **2** (2002) 229−246.
8. H. Xi and E. A. Lunney, *Methods Mol. Biol.* **685** (2011) 279−291.
9. M. Snowden and D. V. Green, *Curr. Opin. Drug Discovery Dev.* **11** (2008) 553−558.
10. S. K. Yeap, R. J. Walley, M. Snarey, W. P. van Hoorn, and J. S. Mason, *J. Chem. Inf. Model.* **47** (2007) 2149−2158.
11. A. S. Bell, J. Bradley, J. R. Everett, M. Knight, J. Loesel, J. Mathias, D. McLoughlin, J. Mills, R. E. Sharp, C. Williams, and T. P. Wood, *Mol. Divers.* **17** (2013) 319−335.
12. T. J. Crisman, J. L. Jenkins, C. N. Parker, W. A. G. Hill, A. Bender, Z. Deng, J. H. Nettles, J. W. Davies, and M. Glick, *J. Biomol. Screen* **12** (2007) 320−327.
13. S. C. Sukuru, J. L. Jenkins, R. E. Beckwith, J. Scheiber, A. Bender, D. Mikhailov, J. W. Davies, and M. Glick, *J. Biomol. Screen* **14** (2009) 690−699.
14. A. S. Bell, J. Bradley, J. R. Everett, J. Howe, J. Loesel, J. Mathias, D. McLoughlin, J. Mills, R. E. Sharp, C. Williams, and H. Zhu, In Preparation (2013).
15. G. A. Bakken, A. S. Bell, M. Boehm, J. R. Everett, R. Gonzales, D. Hepworth, J. L. Klug-McLeod, J. Lanfear, J. Loesel, J. Mathias, and T. P. Wood, *J. Chem. Inf. Model.* **52** (2012) 2937−2949.
16. A. J. Leo and D. Hoekman, *Perspect. Drug Discov.* **18** (2000) 19−38.
17. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, *Adv. Drug Delivery Rev.* **23** (1997) 3−25.
18. H. Wan and A. G. Holmen, *Comb. Chem. and High Throughput Screening* **12** (2009) 315−329.

19. D. Rogers and M. Hahn, *J. Chem. Inf. Model.* **50** (2010) 742−754.
20. SciTegic/Accelrys, Inc.: Pipeline Pilot, San Diego, CA, 2006.
21. R.C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, and J. Smith, *IDrugs* **9** (2006) 199−204.
22. Y. L. Hu, E. Lounkine, and J. Bajorath, *ChemMedChem* **4** (2009) 540−548.
23. R. S. Pearlman and K. M. Smith, *Perspect. Drug Discov.*, (1998) 9−11.
24. Tripos Inc.: DiverseSolutions, St Louis, MO, 2005.
25. D. Weininger, *J. Chem. Inf. Model.* **28** (1988) 31–36.
26. D. Weininger, A. Weininger, and J. L. Weininger, *J. Chem. Inf. Model.* **29** (1989) 97−101.