Bioorganic & Medicinal Chemistry 20 (2012) 5428–5435

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds

Bernd Beck *

Boehringer Ingelheim Pharma GmbH & Co. KG, Department of Lead Identification and Optimization Support, 88397 Biberach, Germany

ARTICLE INFO

Article history:
Available online 20 April 2012

Keywords:
Data mining
Cross reactivities
Frequent Hitter
BioProfile

ABSTRACT

In the last 10–15 years, many new technologies and approaches have been implemented in research in the pharmaceutical industry; these include high-throughput screening or combinatorial chemistry, which result in a rapidly growing amount of biological assay and structural data in the corporate databases. Efficient use of the data from this growing data mountain is a key success factor; 'provide as much knowledge as possible as early as possible and therefore enable research teams to make the best possible decision whenever this decision can be supported by stored data'. Here, an approach which started several years ago to obtain as much information as possible out of historical assay data stored in the corporate database is described. It will be shown how important a careful preprocessing of the stored data is to enhance its information. Different possibilities for accessing and to analyzing the preconditioned data are in place. Some of will be described in the examples.

1. Introduction

For more than a decade, high throughput screening (HTS) has been a common way to start the modern drug discovery process.[1] In addition, the throughput of the follow-up assays has increased dramatically over the last 20 years. Altogether, this has led to an enormous amount of biological assay data points, single dose POC (percent of control) or dose response (DR) values, which are all stored in corporate databases. For some time, the focus has centered on optimizing the data handling from measurement to storage of the data. In this way, the large pharmaceutical companies have produced large and expensive data mountains in their corporate databases.[2] Is the relationship between the costs of generating and maintaining the data and the knowledge we obtain from it for our daily work still balanced? This problem is illustrated in Figure 1.

It appears that generating knowledge from data is more and more becoming the key success factor in many areas. Considering the rapidly growing data mountains in corporate databases this will be an important success factor in the future. Figure 2 shows the increase of the assay data stored in the Boehringer Ingelheim (BI) corporate database within the last decade.

In addition, not only the proprietary data mountain in companies is growing but also the amount of data available from the public domain, such as, for example, the PubChem Bioassay database.[3]

The first publication about efficient use of data as a success factor for pharmaceutical companies was published as early as 10 years ago.[4] Data mining became the buzz phrase in this respect. An apt quotation from the WWW summarizes the value of data mining; 'Data mining is a powerful tool for digging deep into enterprise data to reveal underlying patterns and relationships…'.[5] Although it had been used successfully for structure related properties to define rules like the *rule of five*[6–9] or lead-, drug and fragment likeness[10] data mining in biological assay data was not very common a decade ago. Several years ago, the first papers that used assay data to analyze the behavior of so called frequent hitters[11,12] or promiscuous inhibitors[13,14] and to predict such behaviour[15–17] appeared. Data mining in drug discovery is summarized in more detail in Ref. 18 and 19.

Here, an approach (BioProfile) is described that is used within Boehringer Ingelheim (BI) to generate knowledge from the in-house assay data.

Driven by the question of how to prioritise hit classes from HTS, we wish to determine:

Are there selectivity issues for a given compound/compound class?

Is a given compound a real hit?

A frequent hitter?

An artifact of the assay technology?

Abbreviations: HTS, high throughput screening; BI, Boehringer Ingelheim; POC, percent of control; DB, database; DR, dose response.

* Tel.: +49 (0)7351 548151; fax: +49 (0)7351 838151.
E-mail address: bernd.beck@boehringer-ingelheim.com

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.04.023

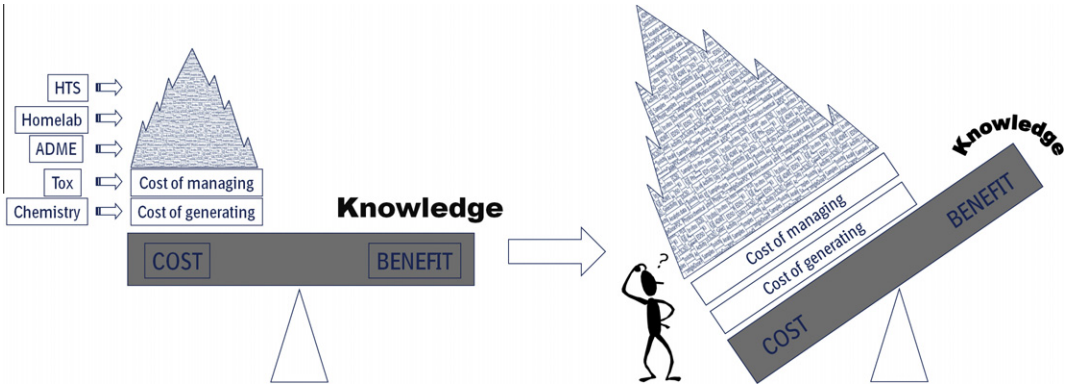


Figure 1. Data mountain—too much of a good thing?

Can we identify selectivity targets not known in advance for the HTS hit set or an interesting hit class?

Are there known toxic effects?

Which data are used, how the data are preprocessed and how we use the resulting information during the project work will be described.

2. Material and methods

The aim was to develop an automated system for the BioProfile analysis. The total workflow consists of three parts; automated and regular data retrieval from the corporate database for all new data, preprocessing of the data, and storage of the preconditioned data in a specialized database.

2.1. Data retrieval

For the analysis of single dose measurements, we currently concentrate on values from HTS screening campaigns and on dose response data from our corporate database. During first trails with the data retrieved a few years ago, we soon discovered that it would be very helpful if we are able to obtain some additional assay specifications from our colleagues from the HTS units within BI global research. This includes information like agonist or antagonist screen, hit threshold, mean value and standard deviation of the screens and also information about the target type and the assay technology. We obtain this additional data directly via DB transfer or via manual import using the KNIME[20] workflow system. All this additional data is again stored in a separate DB. The list currently contains more than 220 primary screens. A weekly update retrieves new single dose data for the stored assays from the corporate database.

For the dose response data we retrieve all IC_{50}, EC_{50}, K_i, K_d, pEC_{50} and pIC_{50} values stored in the central database. Again, an automated weekly update process retrieves new data.

2.2. Data preparation and data storage

Preprocessing of the retrieved data is an essential part of the whole process. Single dose data and dose response data are handled separately.

2.2.1. PrimScreen-profile

Having retrieved all the POC values from the database, we retain the minimum values for antagonistic screens and the maximum ones for agonistic screens if multiple measurements per compound and method exist. The POC value of a compound is needed to check whether it was above or below the hit threshold in the regarded assay. There is no direct comparison of the values between different assays to avoid problems of different screening concentrations.

In this way, we generate approximately 20 million agonist and 150 million antagonist data points. Finally, we add the assay technology and target type information to each assay in the list, as shown in Table 1. The preprocessed data are stored in a separate database instance.

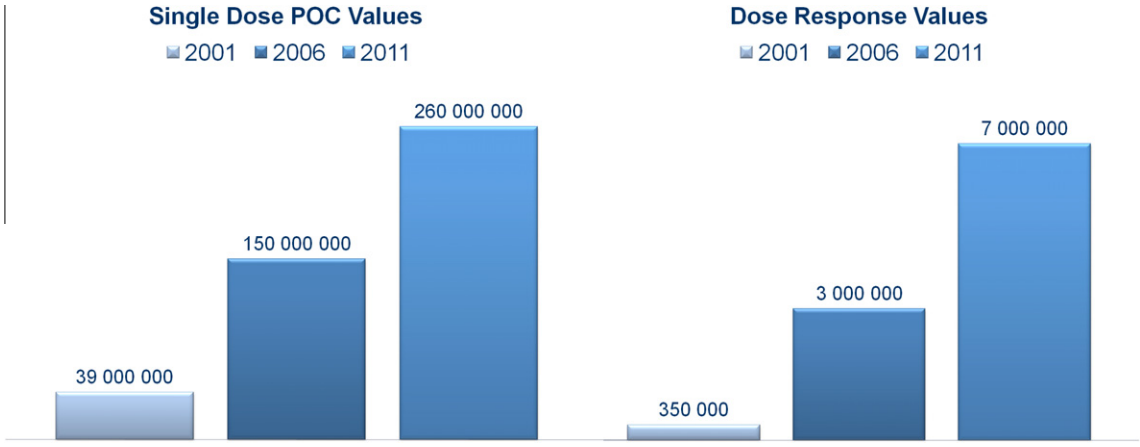


Figure 2. Development of the stored single dose measurements and dose response values in the BI corporate database.

Table 1
Classification of the target and technology types

Assay technology	Target type
Absorption	Enzyme
AlphaScreen[21]	GPCR
Delfia[21]	Ion channel
FlashPlate[21] blue	Kinase
FlashPlate[21] red	Nuclear receptor
FLINT (fluorescence intensity)	Nucleic acid/protein binding
FP (fluorescence polarization)	Phosphatase
FLIPR[23]	Polymerase/nuclease/ helicase
HTRF[22]	Protease
HCS (high-content screening)	Protein binding
LANCE[21]	Transporter
Luminescence	Other
SPA (scintillation proximity assay) blue	
SPA (scintillation proximity assay) red	
FRET (fluorescence resonance energy transfer)	
Other	

2.2.2. Dose response profile

The dose response data from the corporate database also need some preprocessing. First of all, values are converted to μM. For multiple measurements (without an operator) per compound per method, we calculate the median value. This results in more than 4.6 million data points from more than 4000 different assays. Again this data is stored in our specialized DB.

3. Results and discussion

In this chapter we describe some of the possibilities to use and analyze the BioProfile data currently available within BI.

3.1. Direct access via research project related databases

As the BioProfile data is stored in a database, it can be accessed easily from different database frontends or workflow systems. One of the most frequently used ways is via our so-called project databases. These are Oracle based data marts that include all relevant data for a given research project. An example ISIS Form[24] is shown in Figure 3.

This view allows a quick analysis of the results for one compound. In order to analyze the results for complete hit sets or for clusters, we use the KNIME workflow system.[20]

In the following, results based on the primary screen data and dose response data will be shown.

3.2. Analysis based on primary screen data

3.2.1. Frequent hitter analysis

For this purpose we defined a frequent hitter score that depends on the number of screens participated and the number

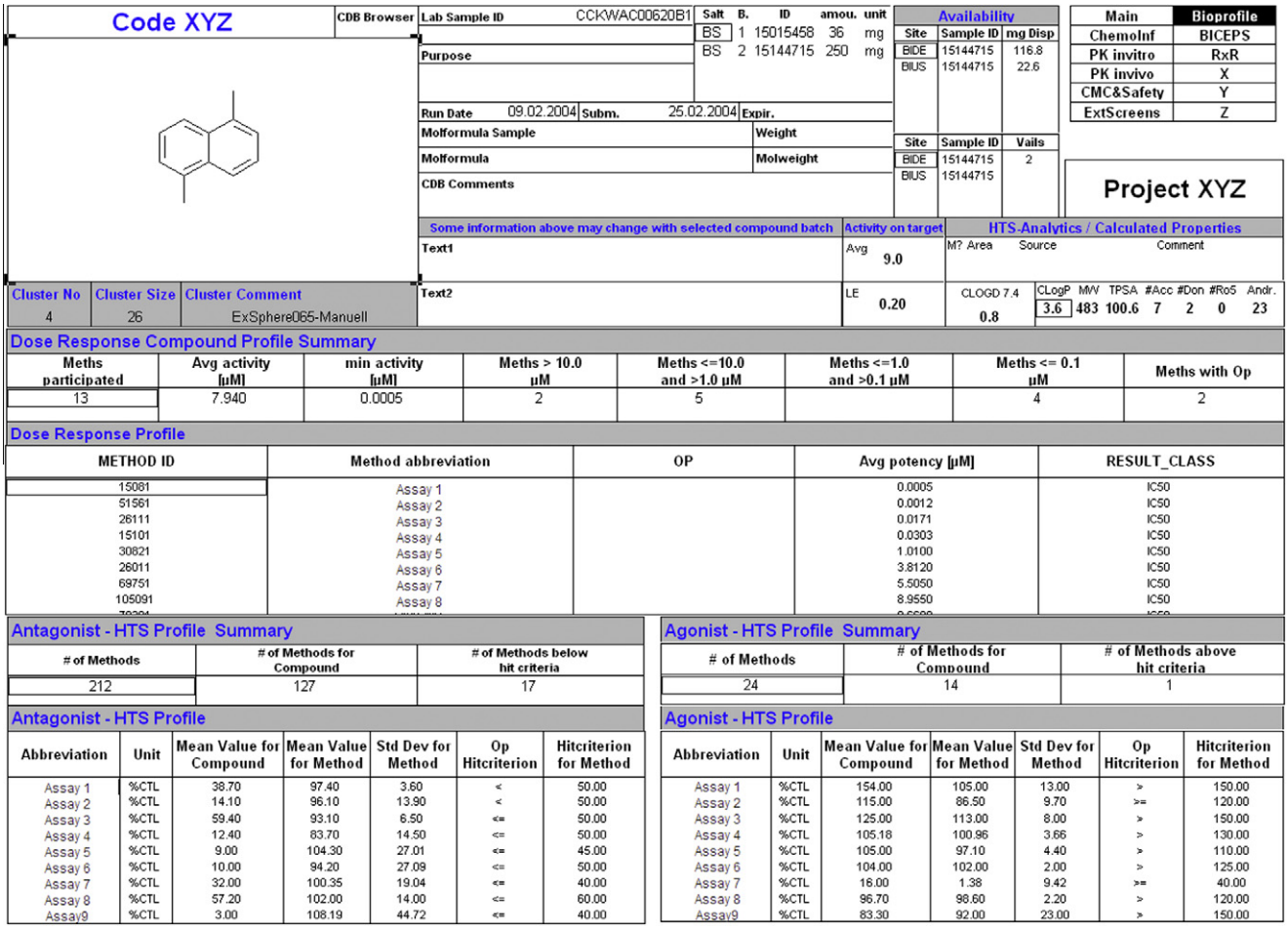


Figure 3. Project DB form with BioProfile data in ISIS base.

Table 2
Example of a frequent hitter analysis result shown for 4 example structures. The list is ranked according to the frequent hitter score.

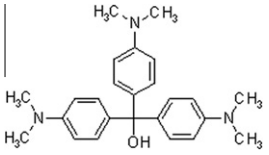
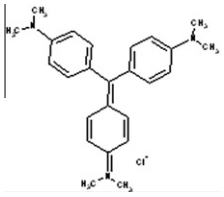
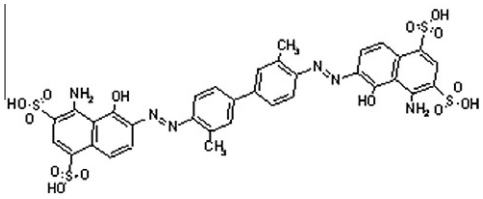
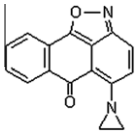
Sample code	Structure	Screens participated	# Screens found as primary hits	Primary hit in # different target types	Primary hit in # different target types	Score	NewScore
ID1		159	78	9	9	214.39	19294.89
ID2		204	79	10	10	193.33	19333.48
ID3		125	56	11	11	147.89	16268.10
ID4		135	56	8	8	142.45	10256.16

Table 3
Overlap of a given hit set with other assays sorted according to the overlap of the actual hit set with the hits from the shown assays

Assay	Assay technology	Assay target type	Hitcrit.	Hitcrit. OP	Hits from hit set	Hits assay	Hit rate assay
Assay 24	AlphaScreen	Protease	50	<	5362	41081	4.2
Assay 21	AlphaScreen	Protein binding	50	<=	3556	99957	10.6
Assay 14	SPA red	Transporter	50	<=	2542	91197	10.5
Assay 1	SPA red	Kinase	50	<	1991	13090	2.1
Assay 100	AlphaScreen	Kinase	50	<	1771	34970	4.3
Assay 34	FLINT	Other	50		1512	31553	3.5
Assay 65	SPA red	Kinase	50	<	1490	10194	1.7
Assay 87	Luminescence	Kinase	50	<	1087	14726	1.7
Assay 213	FlashPlate blue	Kinase	50	<	907	4914	0.7
Assay 17	SPA red	Enzyme	20	<	804	26588	3.8
Assay 33	AlphaScreen	Kinase	50	<	757	15001	2.5
Assay 45	FLINT	Enzyme	50	<	588	8394	0.8
Assay 99	LANCE	Kinase	50	<	526	15410	1.6
Assay156	Luminescence	Protein binding	50	<	492	32521	5.5
Assay 108	AlphaScreen	Protein binding	50	<=	436	14117	1.5
Assay 206	FLIPR	GPCR	50	<	431	43197	4.4
Assay 38	AlphaScreen	GPCR	180	>	196	18759	1.5
Assay 62	LANCE	GPCR	150	>	193	5701	0.6
Assay59	SPA red	Kinase	50	<	187	3465	0.6

of screens where a compound is a primary hit. We aimed at identifying a simple, empirical score that allows us to rank compounds with respect to their promiscuity, also in cases where compounds where tested in a different number of assays. We modeled a biological assay system as a biased coin that yields 'hit' or 'non-hit' with certain probabilities and the various assays to which a compound is subjected as a sequence of independent coin flips. Thus, we use a binomial distribution function to estimate the relative probability of identifying a compound as a hit n times in k independent assays by chance. The probability for the events 'hit' and 'non-hit' were estimated empirically from a set of assays.

The analysis is started by simply joining the precalculated scores with the number of screens in which a compound has participated and the count how often it was a primary hit. An example output with some public domain structures is show in Table 2. For each compound, it is reported how often the compound was in a primary screen campaign, how often it was found as a hit

and in how many different assay technologies and different target types it was found as a hit.

During the discussion of the first results, it became clear that we need an additional frequent hitter score that takes the number of different technologies and target types for which a given compound was found as a primary hit into account. For example, a common kinase inhibitor that was tested in many kinase assays will have a high frequent hitter score even if it was only found as hit in kinase projects. This compound is not a real frequent hitter. We therefore now also use a modified score 'NewScore' that includes this information. Using the NewScore instead of the normal Score to rank the frequent hitters in our database leads to about 200 different structures in the top 1000 list.

A KNIME workflow will now be run regularly to update the frequent hitter information stored in the database. This enables us to check whether compounds in a given hit set are frequent hitters and to mark them with an additional flag. The same analysis can also be performed for each target/technology combination of interest for a research project team.

3.2.2. Analysis of a hit set

In addition to the compound-based analysis of the data, such as the frequent hitter analysis described above, we also use KNIME workflows that allow us to check the overlap of a given hit set with other primary screens. An example of the results obtained is shown in Table 3 and Figure 4.

During the analysis we check how many of the compounds from the current hit set are also found as primary hits in other assays. For all these assays we retrieve the technology and target information, the hit threshold, the number of hits, the hit rate of that particular assay and of course the overlap of the hit sets. All this information allows a fast check of the quality of the actual hits. The appearance of many other assays with the same assay technology with a high overlap of hits might suggest checking for assay artifacts with a technology counterscreen. On the other hand, if many assays with the same target type appear at the top of the list, this is a hint for possible selectivity screens.

This information is visualized in a Spotfire plot[25] in Figure 4. Here, the different combinations of technology and target type are plotted. The number of assays with a specific technology/target combination is shown. The color coding is from green (assays with a low hit rate) to red (assays with a high hit rate) and the size of the circles indicates the overlap of the primary hit sets of a given technology/target combination with the current project hits (sum of overlapping hits). The pies are divided into segments according to the hit rates of the assays included.

The same analysis can also be performed only for the main compound classes of a project, to investigate class specific selectivity or technology effect.

3.3. Analysis based on dose response data

As for the primary screen data there are several possibilities to access and to analyze the data based on DR values. As shown in Figure 3, the easiest way to access the data is via our project DBs. In the following, some additional examples for using the data are shown.

3.3.1. Compound-based profile

As for the primary screen data it is possible to perform a compound-based analysis of the DR data. In this case, the activity in

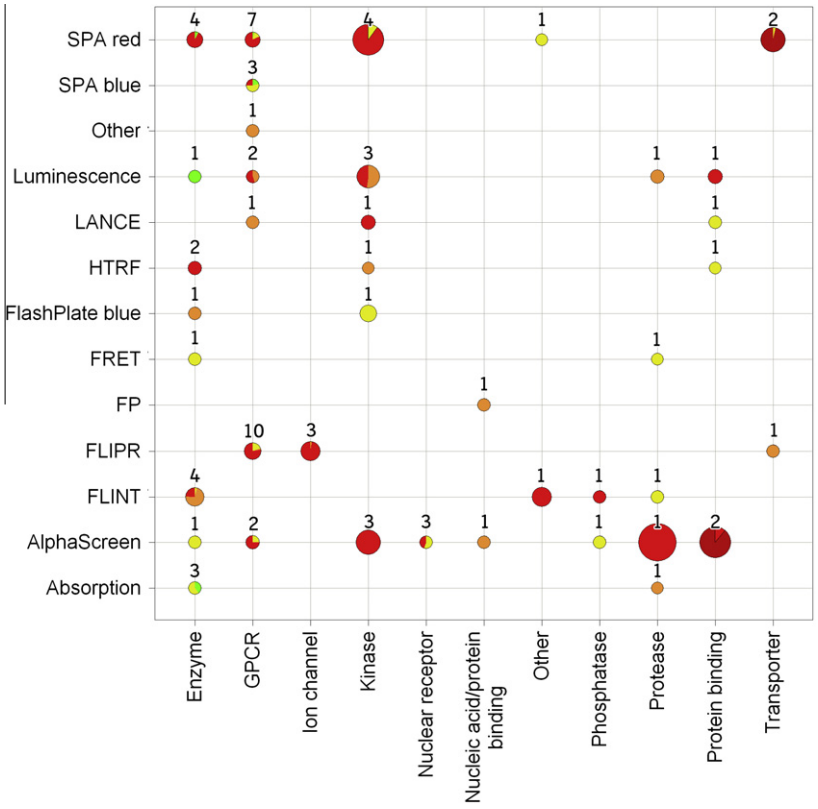


Figure 4. Overlap of the current hit set with other primary screens-—different target/technology combinations are shown. The number indicates the count of assays with a given target/technology combination. The size indicates the overlap of the hit sets of these assays with the current hit list. The color indicates the hit rate of the assays (green: low hit rate–dark red high hit rate).

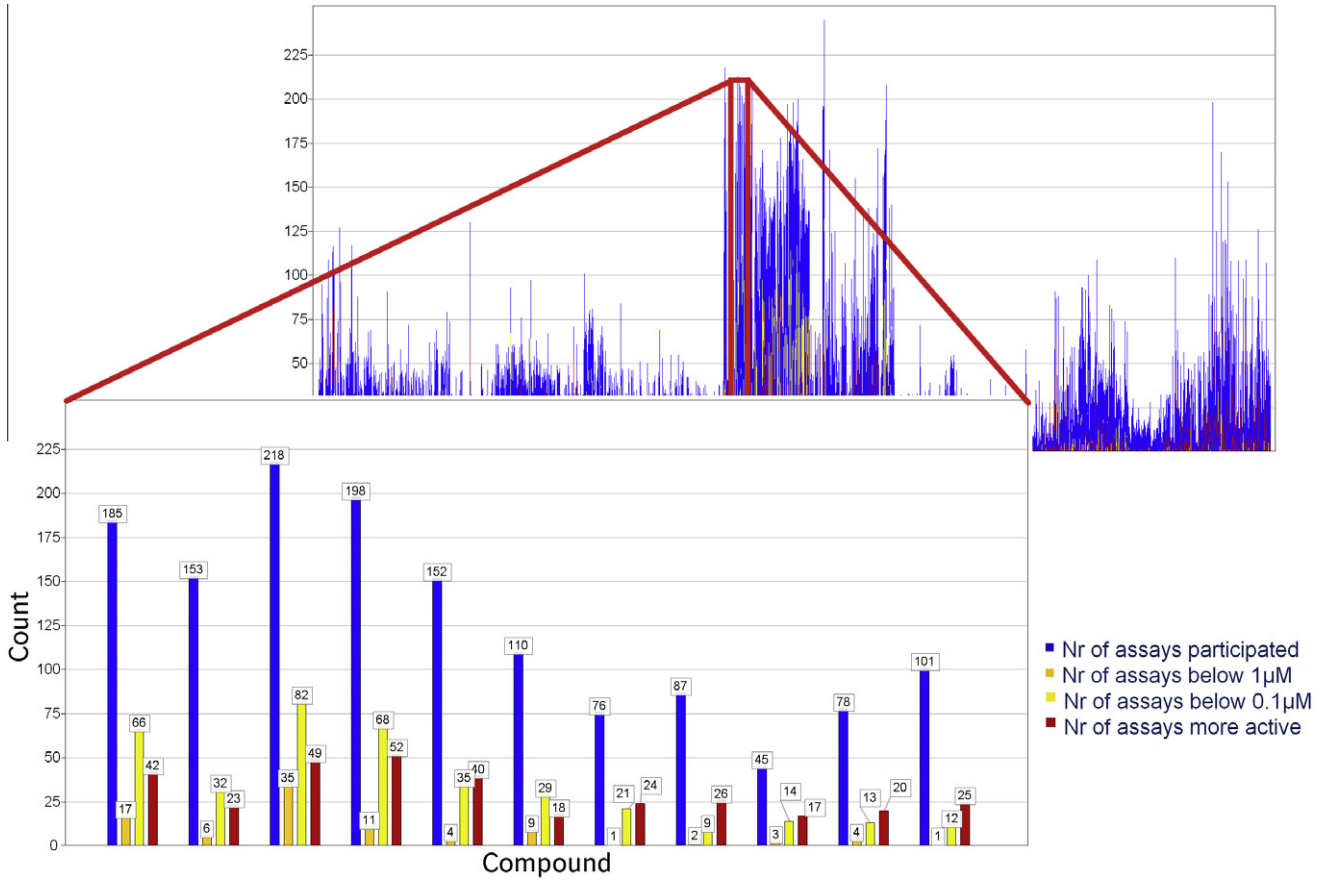


Figure 5. Compound based view based on DR data. In the back the view for the complete hit set analyzed. In the front a detailed view for one compound class is shown.

the various assays is binned. In the example shown, we only count assays in which the activity was below 1 µM and below 0.1 µM. Also shown are the number of tests in which the activity was better than in the current assay and we show the number of assays in which the compounds have a reported DR value in the corporate DB. The values are reported in a table or graphically, as shown in Figure 5. Having this plot it is also possible to zoom in for classes of compounds (see Fig. 5) or for single compounds if necessary. In this way, it is possible to identify cross reactivities for compounds, ADME/TOX issues, like CYP Inhibition or cytotoxicity issues, and with the next version of the DR BioProfile technology issues. The project teams can use this information to prioritize compounds or compound classes.

3.3.2. Hit set profile

In order to gain a quick overview on potential selectivity issues for a given hit set or for a compound class of interest, it is possible to perform an analysis for the complete hit set of a project. Again, the results can be delivered as a table or graphic, as shown in Figure 6.

In Figure 6a the complete plot is shown. Figure 6b shows a zoomed in area of Figure 6a.

In the plot there are regions in which only a few values or only values with '>' operator (not plotted) can be found (e.g., around assay ID 5850). We are not able to draw any conclusions in these areas. On the other hand, there are several assays in which we have a large overlap with the current hit set. This is not a problem if we mostly found green or yellow symbols (e.g., around assay ID 5550) which means that the compounds of the actual hit set are not very active in this assays. If we find many red or dark red compounds for an assay (e.g., assay IDs between 4550 and 4750 or assay ID 4890), we need to check the assay results more closely. Many of the compounds of the given hit set are also active in these targets and it needs to be checked if one or more of these assays should be used as selectivity counterscreens in the current project.

Analyzing the plot in more detail, for example for assay ID 4400 (red and yellow squares), one can already find compounds that are selective against this target and therefore possibly obtain some first hints about how to approach the cross reactivities.

3.4. Analysis of cytotoxicity data

The selection of compounds for a cellular assay serves as the final example for the daily use of the BioProfile in Research. The task was to select hits from the biochemical assay that have an activity below 0.1 µM and show no cytotoxicity effect (>10 µM) in the assays available in the corporate DB. Since the assays are flagged as cytotoxictiy tests in the preconditioned DR BioProfile it is a simple DB query to retrieve the values for the compounds of interest. In the case that several measurements per compound are available, the assay with the lowest DR value is selected. Such an example is shown in Figure 7. All compounds in the green rectangle were selected for the cellular assay.

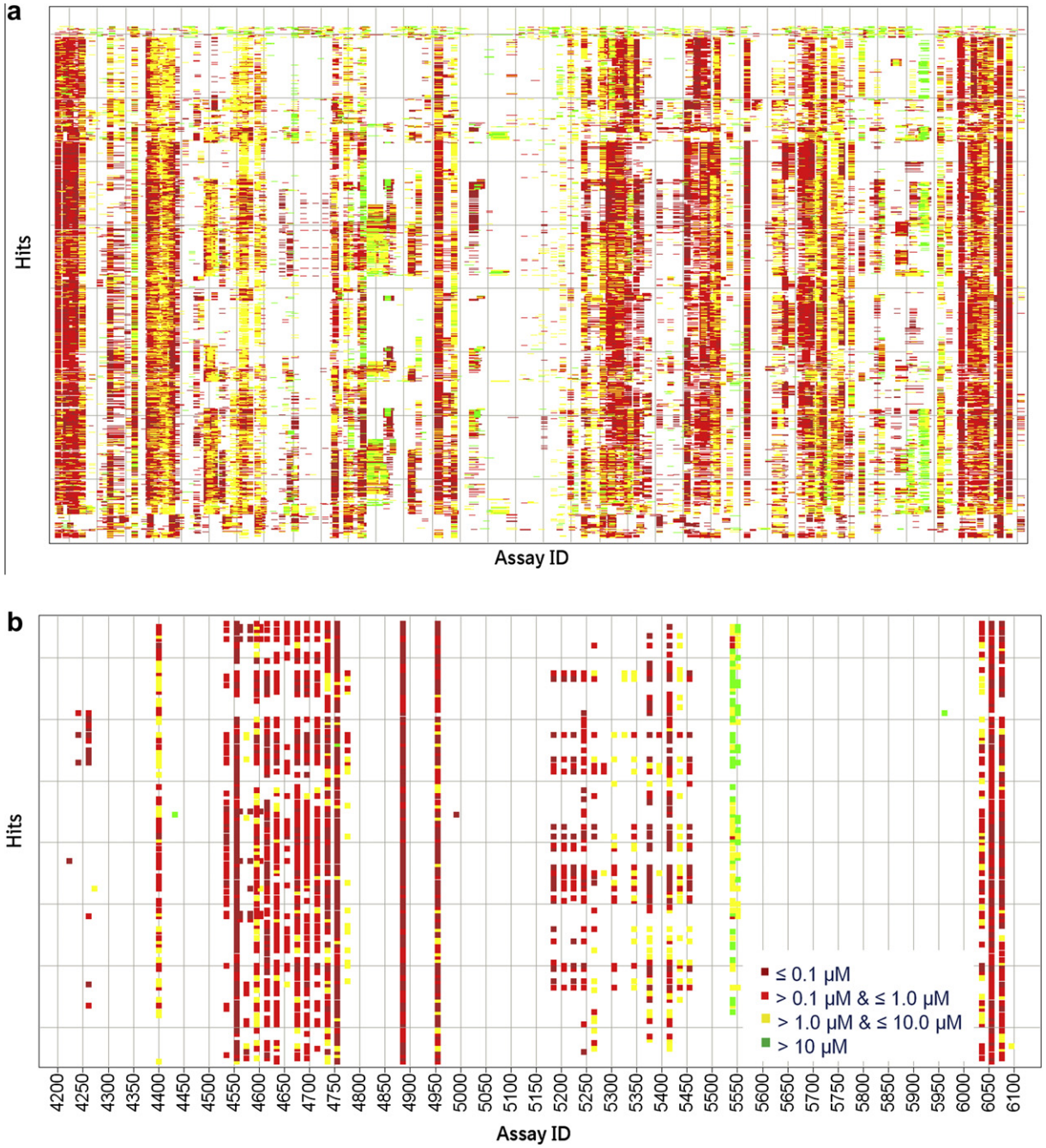


Figure 6. Analysis of the DR data for a compound class or complete hit set; it easily can be analyzed which other assays overlap with the given hits. The color coding is from green (inactive) to dark red (very active) compounds. (a) The complete hit set is shown; (b) zoomed in area of (a).

4. Conclusion

Over the last few years we have implemented a procedure that automatically extracts and preprocesses assay data from the corporate database. In order to increase the information content of the data, we also collect additional information such as the target type and the technology type, from the HTS groups for all screening campaigns. We join this information with the preprocessed assay data and make it all available in a separate DB instance.

There are different ways of accessing the data. Most frequently used is the direct access via our project related databases and via the KNIME workflow tool.

With our BioProfile approach we make it possible for our research project teams to access information about cross reactivities within a given hit set using the assay data stored in our corporate database easily. The data are used to prioritize compounds or compound classes. They are also used to check for selectivity targets or counter assays and to identify assay artefacts. Identification of frequent hitters in the screening collection is another example of use.

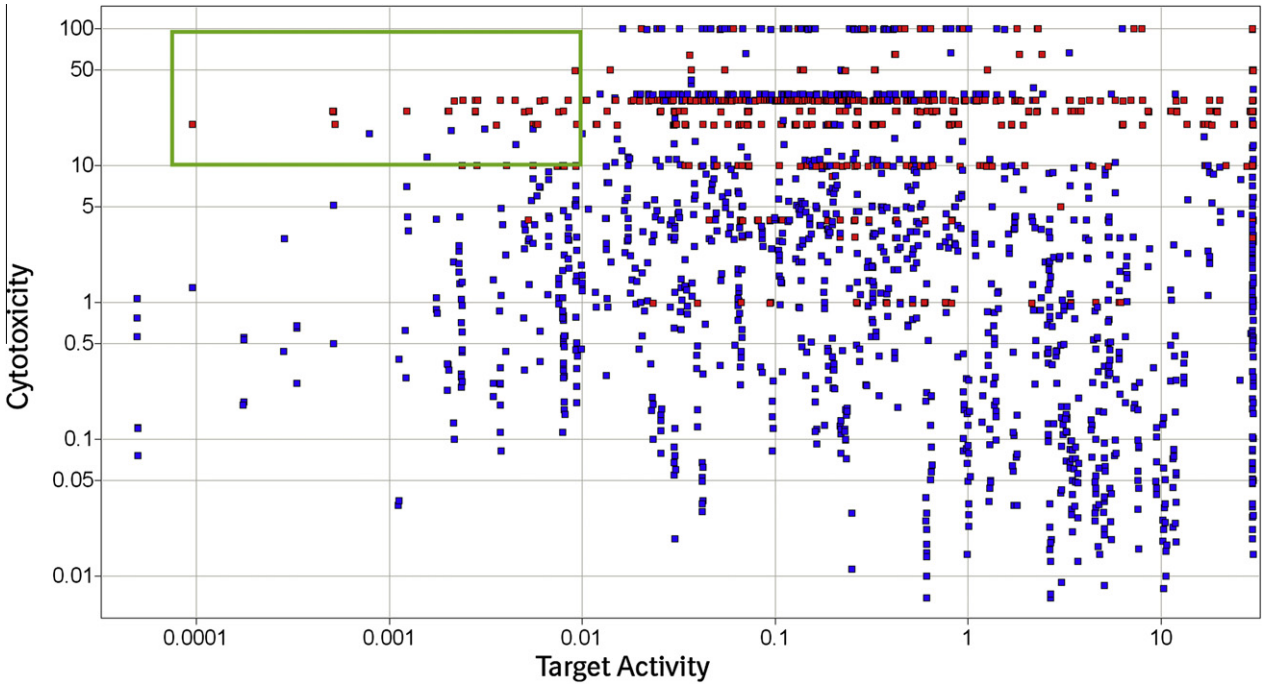


Figure 7. Plot of the activity of the target against cytotoxicity activity. The green rectangle indicates the interesting compounds.

Currently, we are implementing a procedure also to obtain the additional information for all assays with reported dose response in the corporate database. For older assays, this was achieved by manual annotation. For all newer assays, this info is stored in the corporate database. In addition, we are currently including more single dose measurements and therefore more assays from the corporate database in the BioProfile system.

Acknowledgements

The author especially thanks Hans Schubert, a former colleague from the HTS group at our research site in Biberach, who started working with the author on the BioProfile and supported the further development by manually annotating several hundreds of assays with additional technology and target information. Special thanks also to Renate Schnitzer (Vienna) and Richard Nelson (Ridgefield) for providing additional information for primary screens at our research sites in Vienna and Ridgefield.

For the continuous support and fruitful discussions the author thanks the computational chemistry group in Biberach, the HTS group in Biberach and the IS support in Biberach.

References and notes

1. Macarron, R. *Drug Discovery Today* **2006**, *11*, 277.
2. Ekins, S.; Shimada, J.; Chang, C. *Adv. Drug Deliv. Rev.* **2006**, *58*, 1409.
3. Pubchem bioassay database: http://pubchem.ncbi.nlm.nih.gov.
4. Lanfear, J. *Nat. Rev. Drug Disc.* **2002**, *1*, 479.
5. Agheyisi, R. *Mining Data for Business Insights*, 2009. http://mybibeat.wordpress.com/.
6. Navia, M. A.; Chaturvedi, P. R. *Drug Discovery Today* **1996**, *1*, 179.
7. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3.
8. Lipinski, C. A. *Curr. Drug Disc.* **2001**, 17.
9. Lipinski, C. A. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235.
10. Lipinski, C. A. *Annu. Rep. Comput. Chem.* **2005**, *1*, 155.
11. Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. *J. Med. Chem.* **2002**, *45*, 137.
12. Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. *J. Chem. Inf. Model.* **2007**, *47*, 1319.
13. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. *J. Med. Chem.* **2002**, *45*, 1712.
14. Morphy, R.; Rankovic, Z. *Drug Discovery Today* **2007**, *12*, 156.
15. Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. *J. Med. Chem.* **2009**, *52*, 2067.
16. Böcker, A.; Bonneau, P. R.; Edwards, P. J. *J. Biomol. Screen.* **2011**, *16*, 765.
17. Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. *Nat. Biotechnol.* **2006**, *25*, 805.
18. Mucke, H. *Data Mining in Drug Development and Translational Medicine*; Insight Pharma Reports, Cambridge Healthtech Institute, 2009. http://www.insightpharmareports.com/data_mining.
19. *Pharmaceutical Data Mining*; Balakin, K. V., Ed.; John Wiley & Sons: Hoboken, 2010.
20. KNIME: The Konstanz Information Miner. In *Studies in Data Analysis, Machine Learning and Application, Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.*; Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., Eds.; Albert-Ludwigs-Universität Freiburg 207: Springer Heidelberg, 2008; pp 319–326.
21. Registered Trademark of PerkinElmer, Waltham, United States. http://www.perkinelmer.com.
22. Registered Trademark of CisBio Bioassays, IBA group, Louvain-la-Neuve, Belgium. http://www.htrf.com.
23. Registered Trademark of Molecular Devices, LLC, Sunnyvale, United States. http://www.moleculardevices.com.
24. MDL ISIS Base 2.5/SR 3, 2003.
25. Spotfire DecisionSite 9.1.1, Spotfire AB, 2008.