Medicinal Chemistry Tools:
Making sense of HTS data

Evgueni Kolossov*, Glyn Williams; ID Business Solutions Ltd
*Corresponding author; E-mail address: ekolossov@idbs.com (E. Kolossov)

Rational

High throughput screening (HTS) is an expensive part of the drug development process. Increasing the efficiency and productivity of HTS is a key objective for today's discovery organisations. Predictive technology can be used to direct the screening of compounds prior to synthesis. However, issues such as the small proportion (usually less than 1%) of hits that occur in a given assay, complicate the application of statistical analysis and predictive modelling to this data.

What is required is an analysis technique that increases the significance of the data for the active compounds, while using all the information present in the original data set.

Introduction

The recent publication of the genome of the human malaria parasite, *Plasmodium faciparum*, will greatly enhance the drug discovery effort in this area, enabling identification of novel molecular targets. However, small molecule inhibitors that interact with these targets are essential for the development of new antimalarial drugs and the elucidation of the role of newly identified targets.

Erythrocyte invasion by the malaria merozoite requires the activity of parasite serine proteases, and can be prevented by serine protease inhibitors. Compounds which selectively inhibit proteases involved in erythrocyte invasion and other aspects of erythrocytic growth have potential for development as antimalarial drugs. Research aimed at characterizing these proteases led to the identification and recombinant expression of a *P. falciparum* subtilisin-like serine protease called PfSUB-1. The product of a conserved single-copy gene, PfSUB-1 displays relatively high substrate specificity, is expressed in a subset of secretory organelles in the mature blood-stage schizont and merozoite and is implicated in invasion or post-invasion events.

This poster presents a real-world case study carried out between MRCT, the technology transfer company associated with the UK's Medical Research Council, and IDBS. Together, a method was developed for improving the statistical significance of active compounds identified during HTS screening. The resulting 'enhanced' data is used to develop predictive models that can then be used to direct screening programs in real time.

Materials and Methods

This study was based on the screening results for 10,000 compounds against a Malaria PfSub-1 serine protease inhibition assay [1], which measures inhibition of a protease important in the blood stage of the malarial parasite. MRCT had performed an initial round of diversity screening that was only partially successful in identifying hits. None of these 'hit' compounds were sufficiently potent to carry forward.

IDBS used its predictive software [2] to analyze these initial hits in an effort to identify the substructural components that contributed towards the potency of the hits. The objective was to aid the selection of compounds for the next stage of the screening process.

Similar to that of processing sound when the level of noise is higher than the signal, the noise must be reduced without damaging the main signal. Taking this analogy, a method was developed to flatten the frequency distribution of HTS data, based on the activity/property distribution function, see Figure 1. (Full details of the method can be found in [5].)

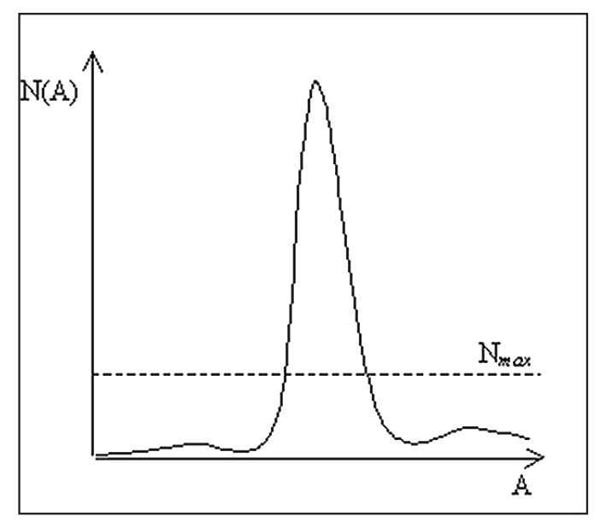


Figure 1

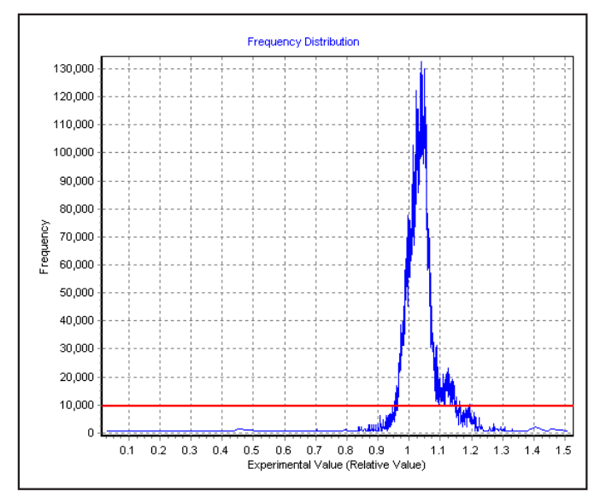


Figure 2

Figure 2 shows the frequency distribution graph for an initial set of 10,000 compounds. The data indicates that the number of the hits in this assay is less than 1% of the total number of compounds.

With this number of active compounds, any statistical analysis will lead to elimination of these compounds as outliers. To make the number of active compounds statistically significant, they should represent over 5% of the total number of compounds. Applying the technique described earlier, the total number of compounds was reduced to 1566 compounds while maintaining the same frequency distribution (see Figure 3).

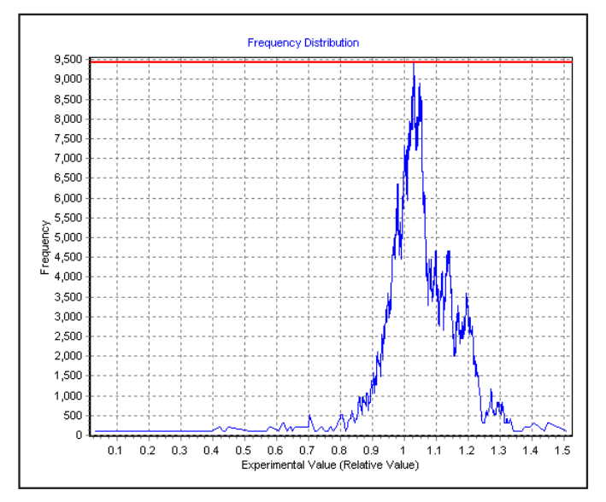


Figure 3

These 1566 compounds were then used to build QSAR models LSF_F(1) and PLS_F(1).

During the fragmentation process, 8716 fragments from the 1566 compounds and 3342 fragments from the 409 compounds were generated from the training set or mapped from a predefined set of fragments. These fragments together with their frequency of occurrence in each structure were used for regression analysis.

Summary statistics for the models are given in Table 1.

Model	r²	Chi²	SE ᵃ	Mean error	Constant	Num. of compounds / outliers	Num. of fragments	Num. of principal components
LSF_F(1)	0.8512	3.877	0.090	0.00228	1.071	1559/7	1080	N/A
PLS_F(1)	0.9414	1.565	0.032	0.00081	0.962	1561/5	8716	21

Table 1. Statistics for generated QSAR models

Figures 4 and 5 show the experimental/calculated graph for PredictionBase LSF_F(1) and PLS_F(1) models respectively, showing the confidence intervals 95%.

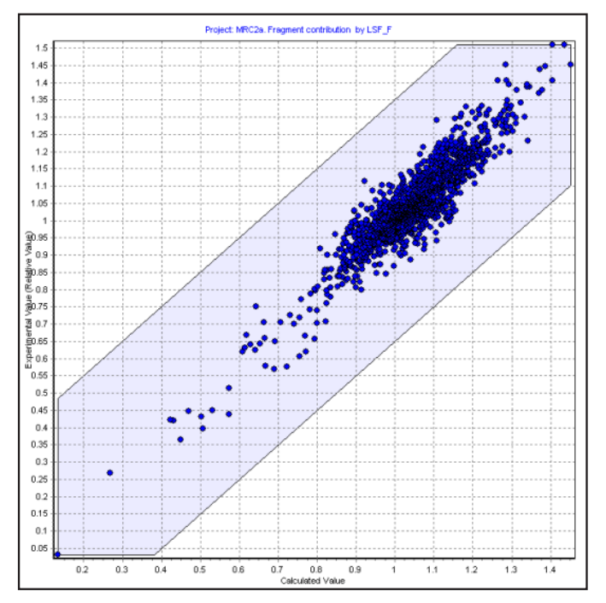


Figure 4

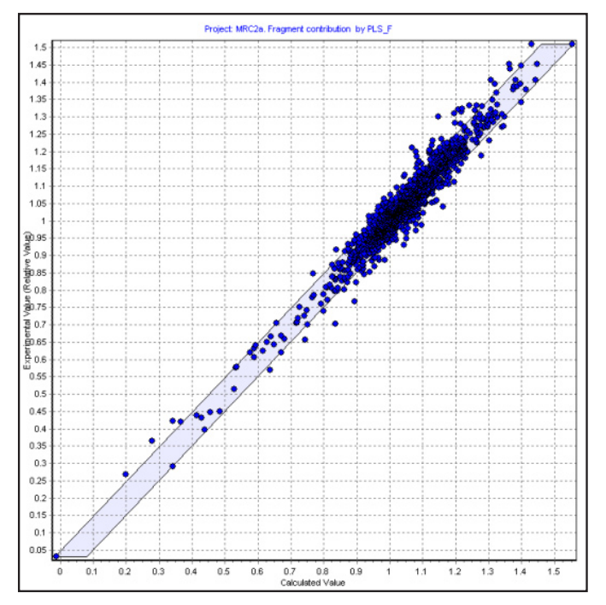


Figure 5

Cross-validation

Leave-One-Out (LOO) cross-validation results are shown in Table 2.

Leave-Many-Out (LMO) validation was performed using the PredictionBase Leave-Group-Out validation procedure, splitting the training set randomly into test groups of 25% of compounds and automatically recalculating new regressions and predictions for this group of compounds. For comparison, the average values from 4 groups (iterations) were used. Comparisons of cross-validation results are shown in Table 2.

Model	q² (LOO)	MIC* (LOO)	r² (LMO)	q² (LMO)	MICᵃ (LMO)
LSF_F(1)	0.8498	2.77	0.9533	-1.8138	120.98
PLS_F(1)	0.9408	197.6	0.9665	0.6996	3841.00

*PredictionBase Model Instability Coefficient

Table 2. Statistics for cross-validation results

Cross-validation results indicate that all the models are relatively stable but become underfitted when 25% of compounds are removed.

To check the possibility of random correlations, the Y-randomization test was performed by scrambling activity values for the whole set of compounds and recalculating regressions. This operation was performed 100 times (iterations). Results are shown in Table 3.

Model	R²minᵃ	r²maxᵇ	Chi²minᵃ	Chi²maxᵇ	SEminᵃ	SEmaxᵇ
LSF_F(1)	0.6461	0.7401	6.771	9.233	0.119	0.139
PLS_F(1)	0.6590	0.7487	6.549	8.886	0.117	0.137

ᵃ Minimum value from the 100 iterations

ᵇ Maximum value from the 100 iterations

Table 3. Statistics for Y-randomization results

Y-randomization test results indicate that the achieved level of random correlation is significantly lower than that of the original regression leading to the conclusion that the models are not random.

External test set validation

A set of 59 compounds was supplied by MRCT without their experimental data as an external test set. Statistics for the test set validation are given in Table 4.

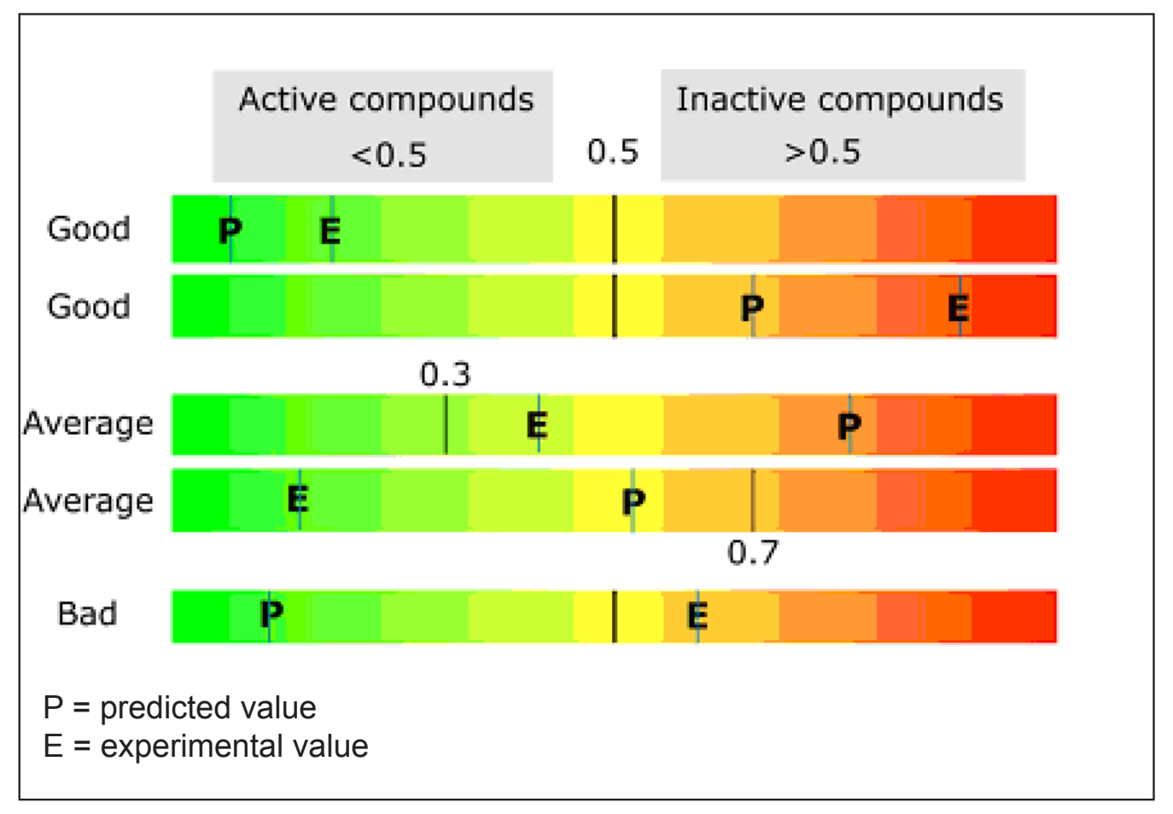
Model	r²	q²	Chi²	SEᵃ	Mean error	Constant
LSF_F(1)	0.8512	-1.763	3.877	0.090	0.0023	1.071
PLS_F(1)	0.9414	0.504	1.565	0.032	0.0008	0.962

ᵃ Standard Error

Table 4. Statistics for external test set validation

To analyze the predictive power of the models, the results have been classified into three groups (summarized in Figure 6 [5]):

1. "Good" results. The activity values have been divided into active compounds (<0.5) and inactive compounds (>0.5). If the compound experimental value was less than 0.5 and the predicted value was less than 0.5, the result is counted as "good". The same applies to the inactive compounds with values over 0.5: if predicted values were over 0.5 it is also counted as a "good" result.
2. "Average" results were classified the same as good results, but with the interval extended to 0.5 ± 0.2.
3. "Bad" results are where the predicted value cannot be referred to as "good" or "average".



P = predicted value
E = experimental value

Figure 6

Model	Num of 'Bad' results	% of 'Bad' results	Num of 'Average' results	% of 'Average' results	Num of 'Good' results	% of 'Good' results
LSF_F(1)	16	27.1	3	5.1	40	67.8
PLS_F(1)	7	11.9	7	11.9	45	76.3

ᵇ Maximum value from the 100 iterations

Table 5. Classification results

These results indicate that the PLS_F model performs better than the LSF_F model and the quality of prediction is improved in models with fewer compounds in the training set. The reason for this is that by applying the flattened distribution method the percentage of active compounds is increased, leading to a corresponding increase in the statistical significance of active compounds.

Conclusions

We have presented a new method for treating HTS data. The results of a Malaria PfSUB-1 serine protease inhibition assay shows that by applying intelligent filtering of HTS data, the statistical significance of the active compounds can be enriched, and generation of predictive models. Flattened distribution filtering is an important development as it broadens the achievability of QSAR techniques to support drug discovery. These models provide medicinal chemists with a powerful tool for optimizing compounds and mining screening candidates in libraries.

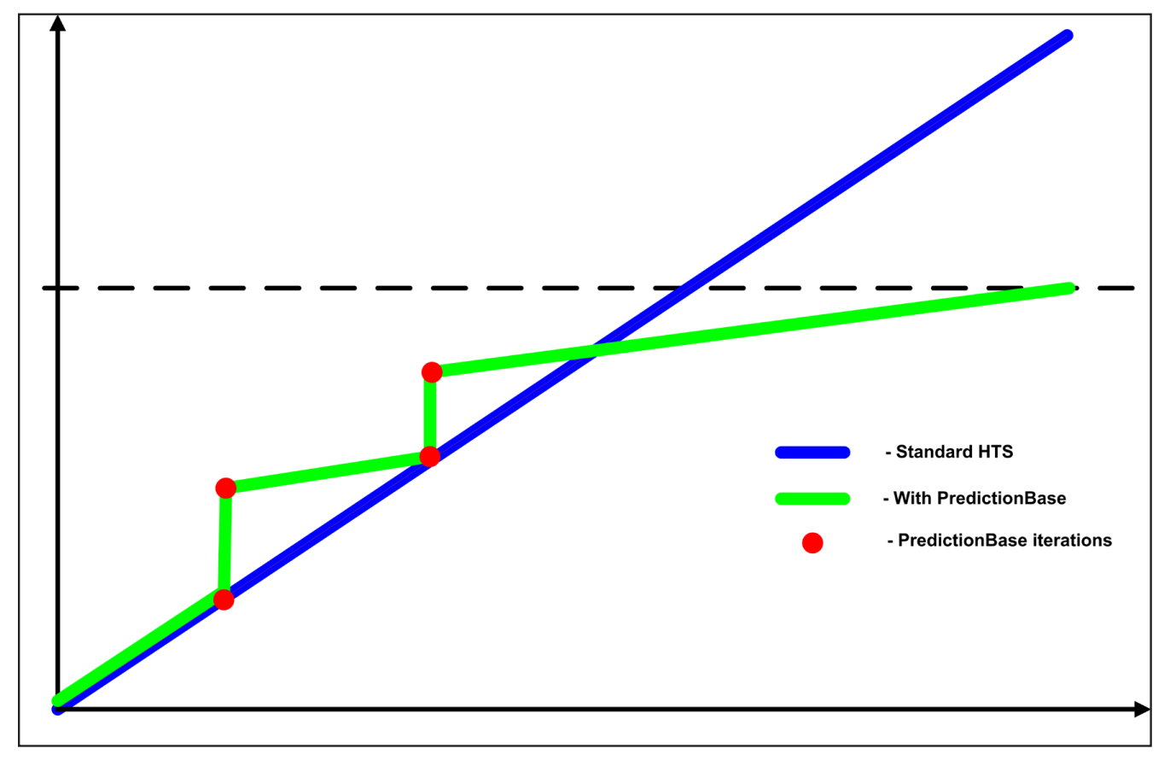


Figure 7 Cost Efficiency

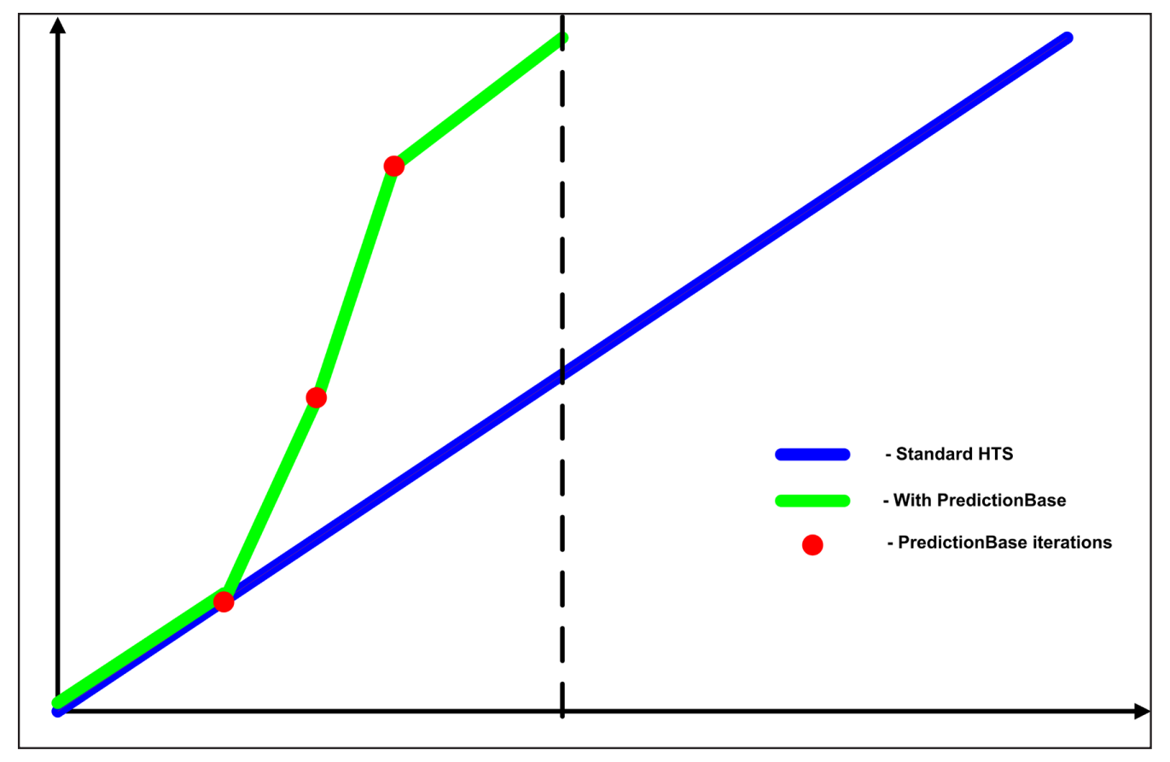


Figure 8 Productivity

The poster presents a simple way of using QSAR models generated in PredictionBase to virtually screen HTS data and increase its productivity. A typical screening campaign can be modified by splitting into the following stages (see Figure 7):

1. Screen just 20% of the total number of compounds.
2. Filter the results using the above discussed technology.
3. Build a QSAR model.
4. Apply the QSAR model virtual screening process to filter all remaining compounds.
5. Repeat step 1 – 4 (with the QSAR model amendment based on new results) until very little to zero hits are found in the virtual screening.

Using this technology the total number of compounds to screen will be reduced by at least a half. Despite the cost of modelling and time spent on virtual screening, the total cost of the screening campaign is also reduced by at least 40% (see Figure 8).

Summary

Predictive models can be generated that intelligently filter HTS data in real time to help guide the screening process. The maximum benefit is gained when the scientists use these predictive approaches directly. By adopting an iterative process of synthesis and testing, and feeding the results back into predictive models, bench scientists have a powerful tool that can be used to mine screening candidates in libraries and aid compound optimization. Empowering these scientists with such predictive capabilities not only contributes to increased productivity and efficiency, but also helps reduce discovery costs.

Acknowledgements

With thanks to Debra Taylor, Keith Ansell, Mike Blackman and Barbara Saxby from the MRC Technology Assay Development group.

References

[1] K. Ansell, B. Saxty, C. Kettleborough, M. Dalrymple, J. Corrie, M. Blackman, Poster presentation: Society for Biomolecular Screening 8th Annual Conference, The Hague, The Netherlands, 22-26 Sep 2002.
[2] PredictionBase 2.0. ID Business Solutions Ltd., 2 Occam Court, Occam Road, Surrey Research Park, Guildford, Surrey, GU2 7QB,
[3] D.M.Hawkins, J.Chem.Inf.Comput.Sci. 2004, 44, 1–12.
[4] W.H.Press, et al., Numerical Recipes in C (2nd edition), 1992, Cambridge University Press
[5] E.Kolossov, A.Lemon, Medicinal chemistry tools: making sense of HTS data, Eur.J.Med.Chem, 2005 (in Press)

One integrated framework for all your discovery data

ID Business Solutions Ltd
2 Occam Court
Surrey Research Park
Guildford, Surrey
GU2 7QB UK
+44 1483 595000

ID Business Solutions Inc
1301 Marina Village Parkway
Suite 320
Alameda, CA 94501
USA
+1 510 8144900

ID Business Solutions Inc
750 US Highway 202
Suite 200
Bridgewater, NJ 08807
USA
+1 908 429 2900

ID Business Solutions Inc
215 First Street
Suite 209
Cambridge, MA 02142
USA
+1 617 995 4422

www.idbs.com