Available online at www.sciencedirect.com
SCIENCE DIRECT
European Journal of Medicinal Chemistry 41 (2006) 166–175
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://france.elsevier.com/direct/ejmech

Original article

Medicinal chemistry tools: making sense of HTS data

Evgueni Kolossov *, Andrew Lemon

ID Business Solutions Ltd., 2 Occam Court, Occam Road, Surrey Research Park, Guildford, Surrey GU2 7QB, United Kingdom

Received 3 May 2005; received in revised form 30 September 2005; accepted 6 October 2005
Available online 20 December 2005

Abstract

The main problem in QSAR modeling from the high throughput screening (HTS) data is that by definition, it produces only a small proportion of hits against a given assay. This leads to a very small statistical significance of the hits in comparison with the "noise". Analysis based purely on the "hit" compounds removes useful information about the biological response of all the test compounds. What is needed is an analysis technique that increases the significance of the active compounds, while using the information present in the original data. In this paper we present a method for application of intelligent filtering of the data to improve statistical significance of the active compounds to generate predictive models that provide medicinal chemists with a powerful tool for both optimizing compounds and mining screening candidates in libraries.

Keywords: High throughput screening; HTS; Statistical significance; Filtering; Optimization; Frequency distribution

1. Introduction

A novel method of analyzing high throughput screening (HTS) data was applied to the results of a Malaria PfSub-1 serine protease inhibition assay, as part of a joint research study with the Medical Research Council Technology (MRCT). Dealing with the large volumes of data generated by HTS is a real issue in drug discovery. Most HTS campaigns produce, by definition, a small proportion of hits against a given assay. Analysis based purely on the "hit" compounds removes useful information about the biological response of all the test compounds. What is needed is an analysis technique that increases the significance of the active compounds, while using the information present in the original data. By applying intelligent filtering of the data, the statistical significance of the active compounds were enriched and used to generate predictive models that provide medicinal chemists with a powerful tool for both optimizing compounds and mining screening candidates in libraries.

2. Methods and data

2.1. Flattening the frequency distribution

In many cases, HTS data are the only data available for building QSAR models. The advantage of this approach is that the data are produced using a highly standardized protocol and the results should therefore be of good quality. A fundamental part of the approach is to screen large volumes of compounds exploring for actives. Unfortunately, there is only a very small chance that these data will contain a significant number of active compounds. In other words, the noise will be higher than the useful information. The problem is that the overwhelming statistical significance of the noise suppresses the useful data and does not allow standard QSAR models to be built. To find a satisfactory solution, methods that are capable of decreasing the noise without affecting the significant data are required.

This problem is very similar to that of processing sound when the level of noise is higher than the signal. Based on this analogy we have developed a method to flatten the frequency distribution of HTS data, based on the activity/property distribution function (see Fig. 1).

The frequency distribution (activity distribution) in activity space $N(A)$ can be defined such that $N(A)\delta A$ is the number of

* Corresponding author. Tel.: +44 1483 59 5000; fax: +44 1483 59 5001.
E-mail address: ekolossov@idbs.com (E. Kolossov).

0223-5234/$ - see front matter © 2005 Elsevier SAS. All rights reserved.
doi:10.1016/j.ejmech.2005.10.005

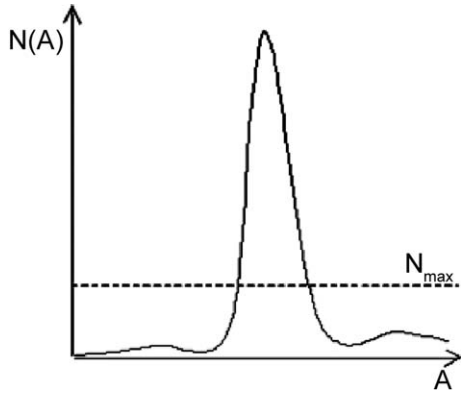

Fig. 1. Activity/property distribution function.

compounds in the range of activity $(A - \delta A/2,\ A + \delta A/2)$. Clearly, $N(A)$ satisfies the criterion:

$$\int_{A_{\min}}^{A_{\max}} N(A)dA = N_{total}$$

where N_{total} is the total number of compounds in the training set. The activity interval δA can be defined as:

$$\delta A = B\Delta A/(N_{\text{total}} - 1)$$

where ΔA is the activity range of the whole training set, and B is the "bin size"—the average number of compounds per bin in the calculation of the activity distribution.

The goal in flattening the distribution is to prevent $N(A)$ from exceeding a set threshold $N_{\max}$, thereby eliminating peaks in the distribution graph. In any region for which $N(A) > N_{\max}$, this is achieved by removing a proportion of compounds to reduce $N(A)$ towards $N_{\max}$.

In an "overpopulated" region of activity space—one in which $N(A) > N_{\max}$—the proportion of compounds considered surplus is:

$$\frac{N(A) - N_{\max}}{N(A)}$$

in the sense that removing this proportion of compounds would yield an activity distribution of $N_{\max}$. In practice it is not necessarily desirable to go to this extreme, and instead some proportion α $(0 < \alpha < 1)$ of the surplus compounds are removed. Although this leaves some regions with remaining activity distribution greater than $N_{\max}$, it better preserves the original shape of the distribution graph.

The choice of which compounds to remove from overpopulated regions is based on random selection. Any compound C (with activity value A_C) is removed with probability:

$$\alpha\frac{N(A_C) - N_{\max}}{N(A_C)}$$

independently of all other compounds.

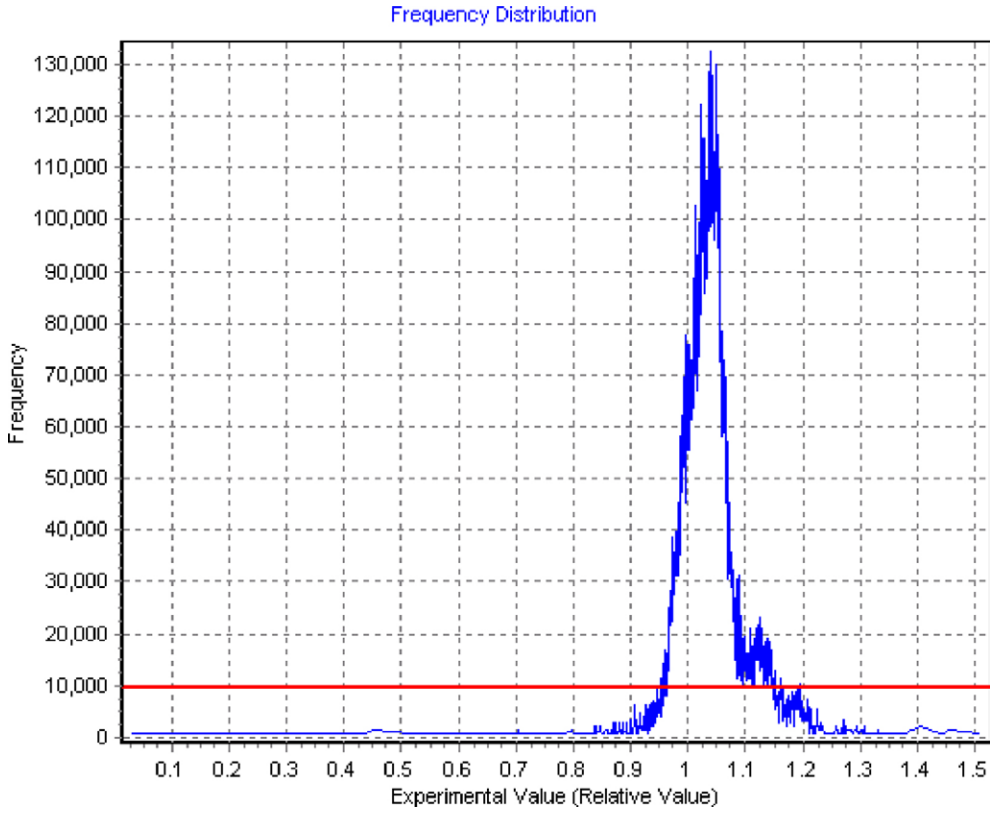

Fig. 2. The frequency distribution graph for an initial set of 10,000 compounds.

2.2. Data

This study was conducted in collaboration with the MRCT based on the 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.

2.3. Software: PredictionBase [2]

The standard PredictionBase algorithm was used to automatically generate fragments from the training set. The following options were selected:

- minimal length of hydrocarbon chain = 4;
- maximal length of hydrocarbon chain = 12;
- use double and triple bonds as fragment centers;
- use full charges on atoms;
- use cycles, aromatic and heterocycles as fragment centers;
- use first and second level of atoms around the fragment centers to generate new fragments;
- not use predefined fragments.

2.4. Correlations

Two different methods were used to generate linear statistical models: least squares fitting (LSF) with stepwise regression and partial least squares (PLS). Both methods are an integral part of the PredictionBase software.

2.5. Confidence region

According to Hawkins [3] variance of each experimental activity σ^2 is assumed constant but is unknown. Standard deviation is square root of variance. Variance of calculated activity for a new structure X is $\sigma^2 h(X)$ where $h(X)$ is the leverage of X [3].

Based on this, the variance of the difference between experimental and calculated activity values for a new structure X is $\sigma^2(1 + h(X))$ and for a structure from the training set is $\sigma^2(1 - h(X))$.

If the number of principal components (variables) is much less than number of structures, then the leverage value is very small [3]. Therefore we can make the simplifying assumption that the variance of the difference between experimental and calculated activity values is equal σ^2 for a new and training structure alike. 95% confidence region (on the experimental/calculated activity graphs) is those points with less than 5% probability of experimental activity value being at least as far as it is from calculated activity values. This is $c\sigma$, where c is 95% (two-tailed) critical point on the standardized normal distribution.

Recall that σ is unknown and has to be estimated. This can be done using the following unbiased estimator [4]: σ^2 = S.S.E./D.O.F., where the sum-of-squared-errors (S.S.E.) is the sum (over training set) of squared differences between experimental and calculated activity values, and degrees-of-freedom (D.O.F.) is the number of structures in training set minus (1 + number of descriptors).

3. Results and discussion

Fig. 2 shows the frequency distribution graph for an initial set of 10,000 compounds.

The experimental data presented in Fig. 2 indicate that the number of the hits in this assay is less than 1% of the total number of compounds. The highest frequency experimental results concern inactive compounds. With the number of active compounds less than 1%, 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 Fig. 3). These 1566 compounds were then used to build QSAR models LSF_F(1) and PLS_F(1). During the second iteration the proportion of active compounds was increased to 25% leaving 409 compounds in the training set used to build QSAR models LSF_F(2) and PLS_F(2) (see Fig. 4).

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.

Figs. 5 and 6 show the experimental/calculated graph for PredictionBase LSF_F(1) and PLS_F(1) models, correspondingly. Figs. 7 and 8 show the experimental/calculated graph for PredictionBase LSF_F(2) and PLS_F(2) models, correspondingly.

3.1. 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 four groups (iterations) were used.

Comparisons of cross-validation results are shown in Table 2.

Cross-validation results indicate that all the models are relatively stable as such but lose stability 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.

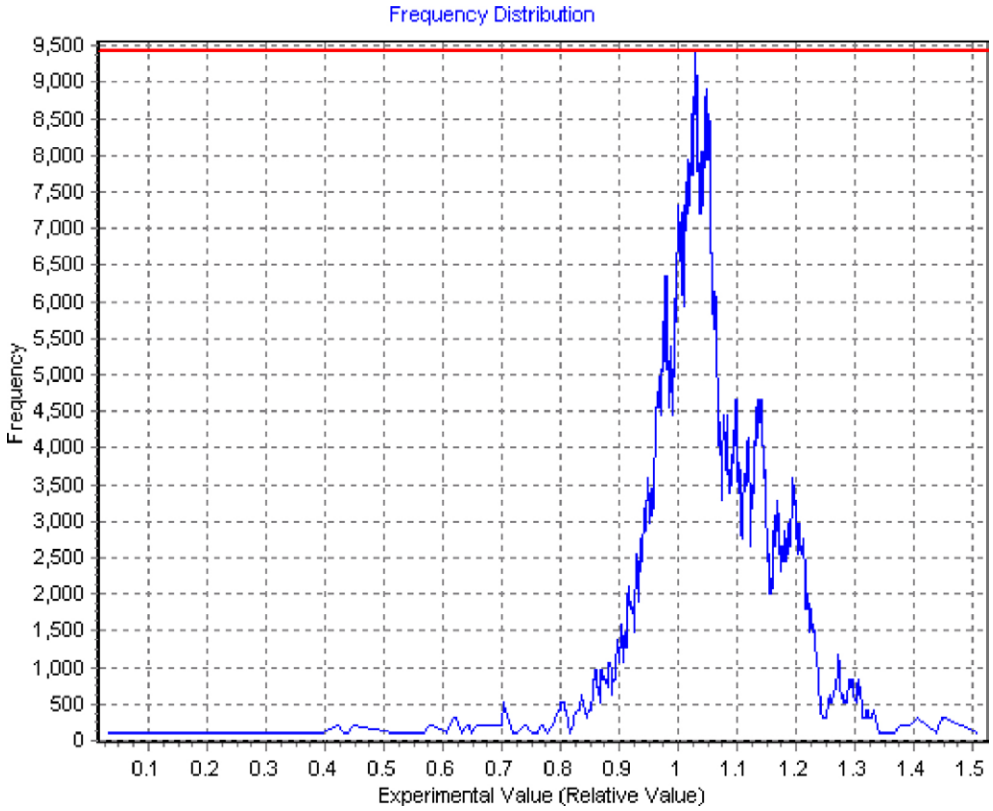

Fig. 3. The frequency distribution graph for the reduced set of 1566 compounds.

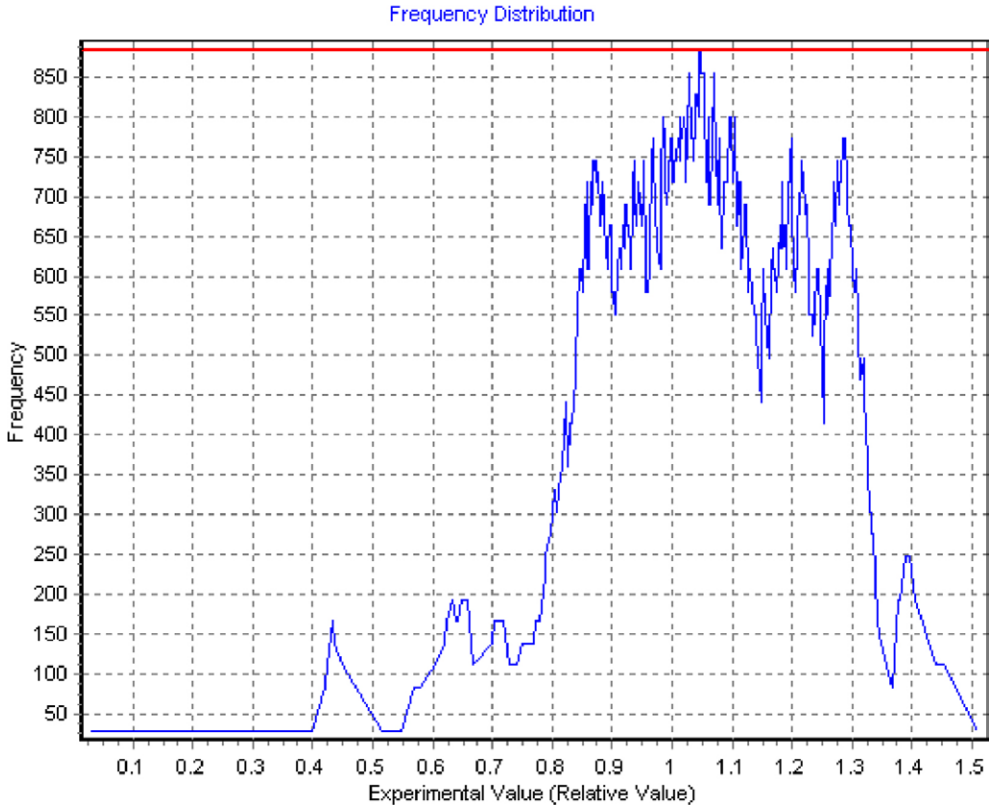

Fig. 4. The frequency distribution graph for the reduced set of 409 compounds.

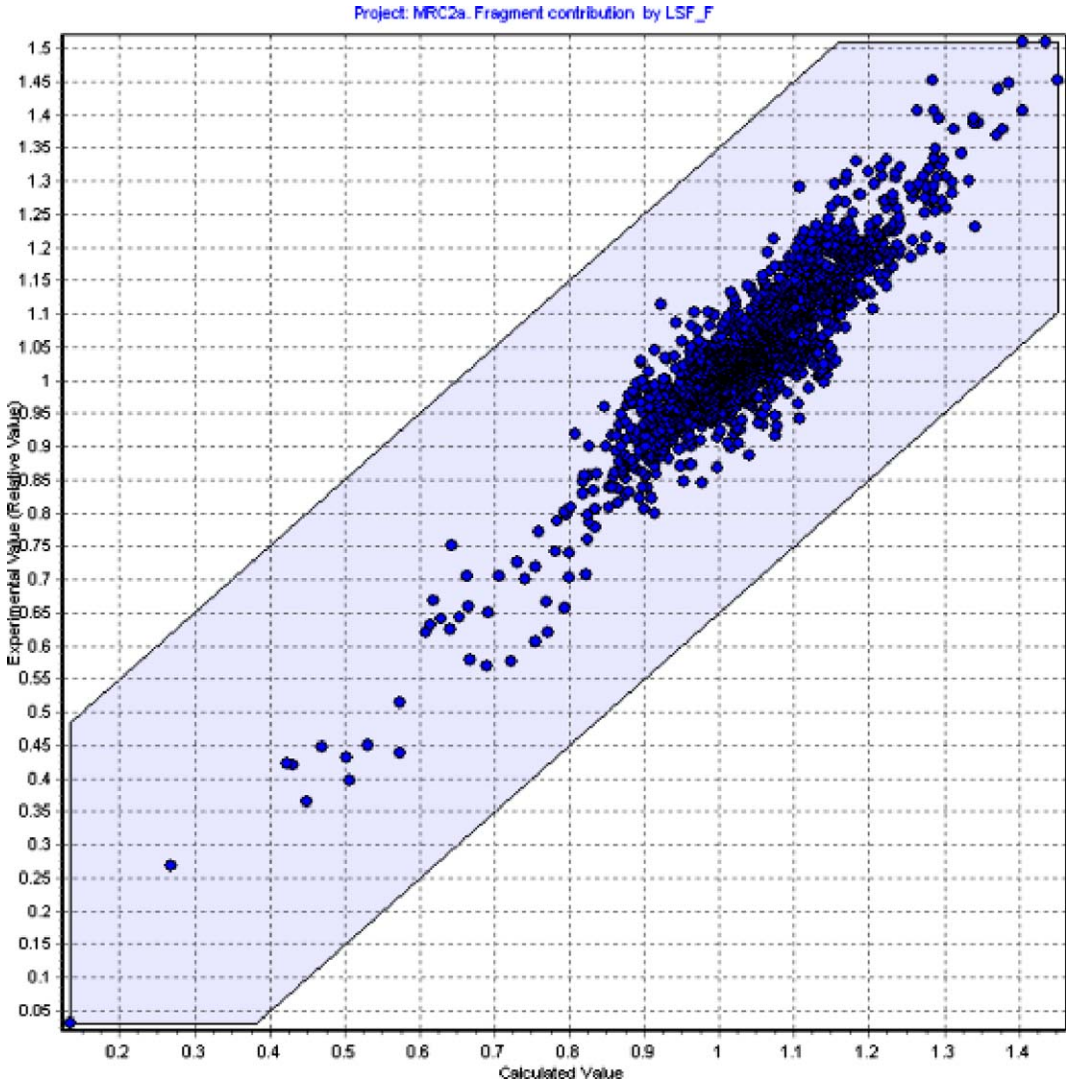

Fig. 5. Experimental/calculated value graph for LSF_F(1) model (blue area is the 95% confidence region).

This operation was performed 100 times (iterations). Results are shown in Table 3.

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. However, the PLS_F(2) model has a relatively high chance of random correlation. This can be an indication that model becomes underfitted after the significant reduction in the number of compounds in the training set.

3.2. External test set validation

A set of 59 compounds was supplied by MRCT without their experimental data as an external test set. The results of this "blind" test are given in Table 4.

Statistics for the test set validation are given in Table 5.

To analyze the predictive power of the models, results have been classified into three groups (summarized in Fig. 9):

- "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 predicted value was less than 0.5, this result is counted as a "good". The same concerns the inactive compounds with value over 0.5: if predicted values were over 0.5 also it is counted as a "good" result.
- "Average" results are those where in comparison with good results the interval is extended to 0.5 ± 0.2.
- "Bad" results are those where the predicted value cannot be referred to as "good" or "average".

Results of this analysis are presented in Table 6.

These results indicate that the PLS_F models perform better than the LSF_F models 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 lead-

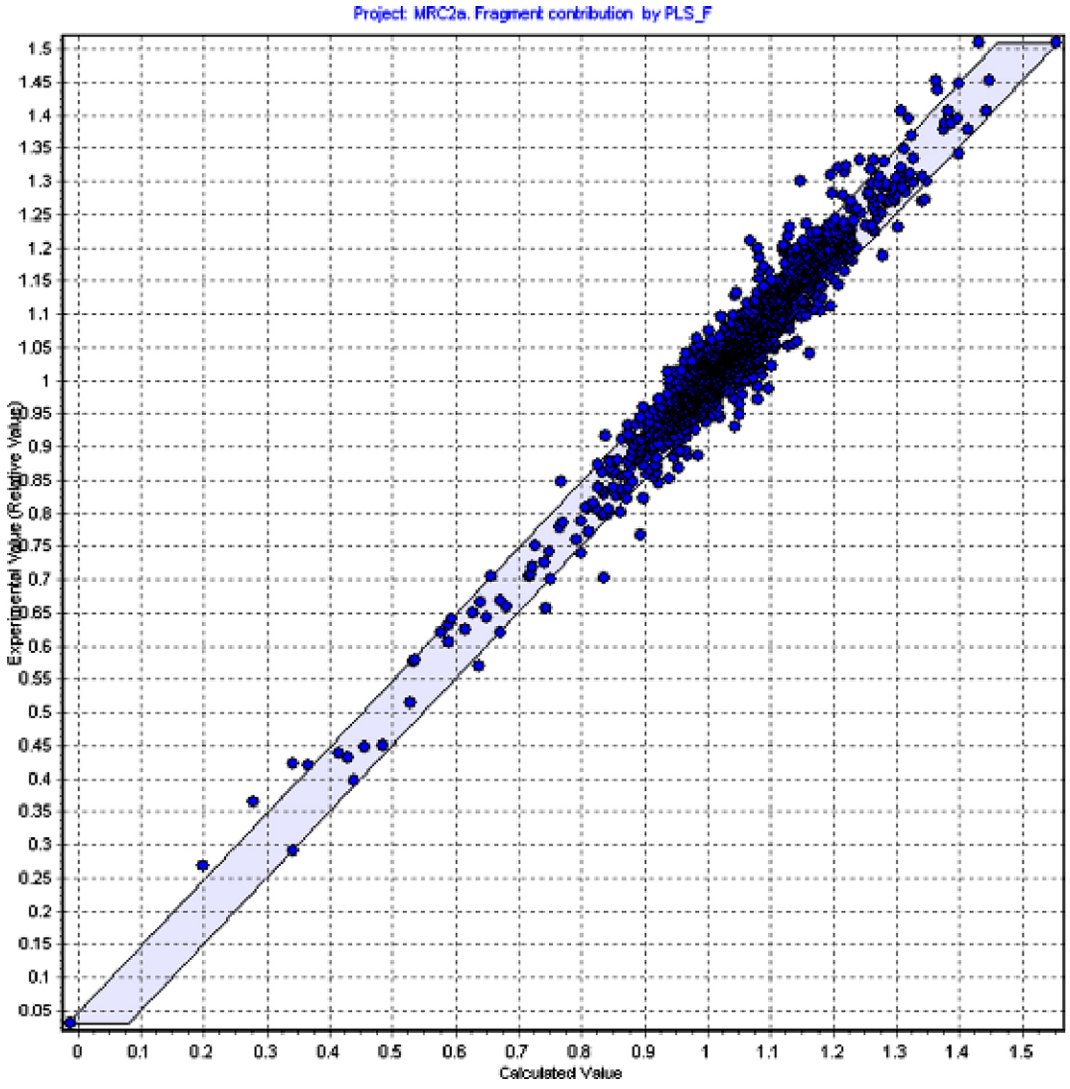

Fig. 6. Experimental/calculated value graph for PLS_F(1) model (blue area is the 95% confidence region).

Table 1
Statistics for generated QSAR models

Model	r^2	Chi square	S.E. [a]	Mean error	Constant	Number of compounds/outliers	Number of fragments	Number 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
LSF_F(2)	0.7648	4.615	0.175	0.00872	1.084	405/4	254	*N/A*
PLS_F(2)	0.9771	0.465	0.034	0.00170	0.757	407/2	3342	12

[a] Standard error.

ing to a corresponding increase in the statistical significance of active compounds.

4. 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, generating predictive models. The application of flattened distribution filtering is an important development in the application of QSAR techniques in drug discovery and will increase the number of problems that can be tackled. These models provide medicinal chemists with a powerful tool for optimizing compounds and mining screening candidates in libraries.

Acknowledgments

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

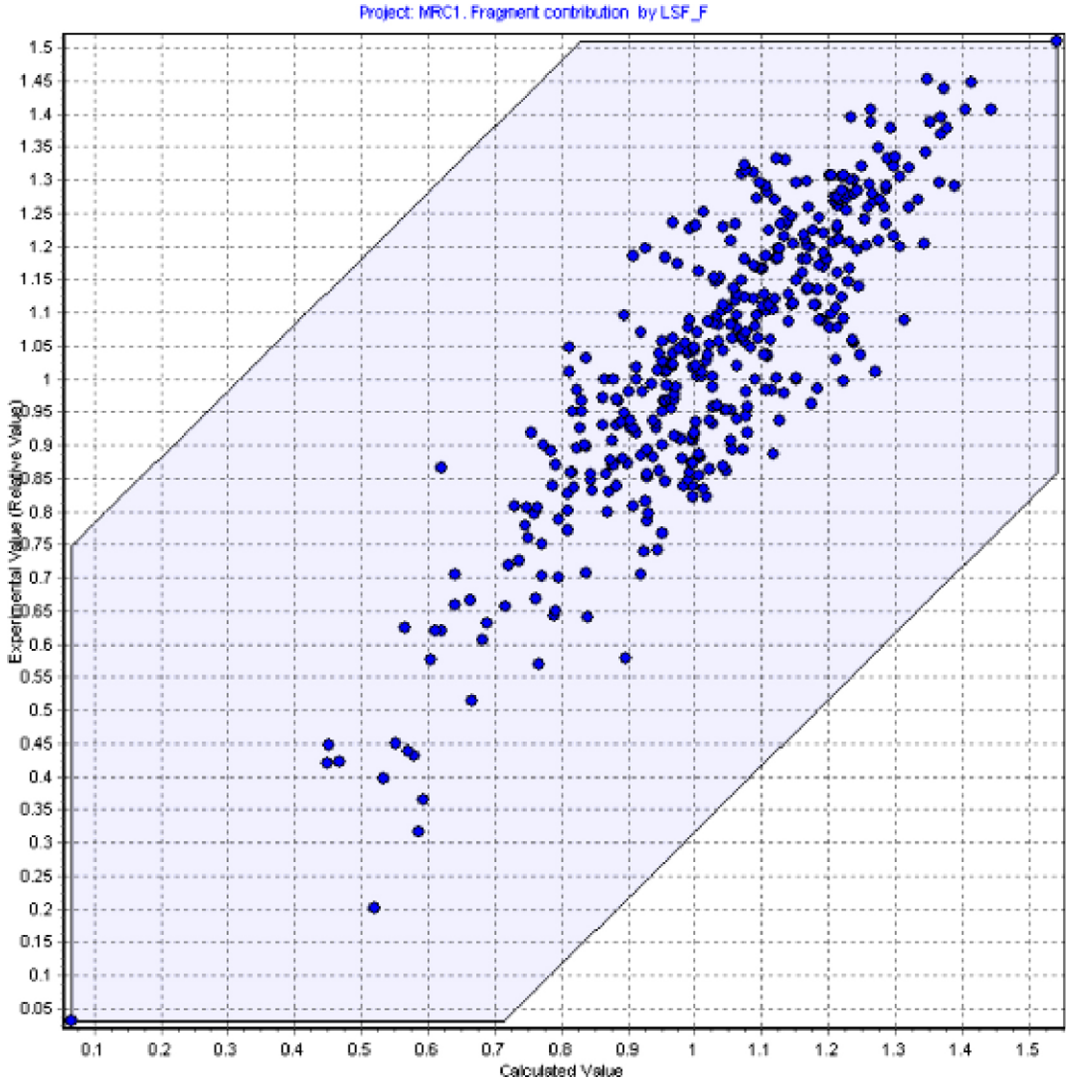

Fig. 7. Experimental/calculated value graph for LSF_F(2) model (blue area is the 95% confidence region).

Table 2
Statistics for cross-validation results

Model	q^2 (LOO)	MIC* (LOO)	r^2 (LMO)	q^2 (LMO)	MIC [a] (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
LSF_F(2)	0.7648	5.25	0.8784	–0.2962	64.63
PLS_F(2)	0.9771	327.4	0.9853	0.6932	3130

[a] PredictionBase model instability coefficient.

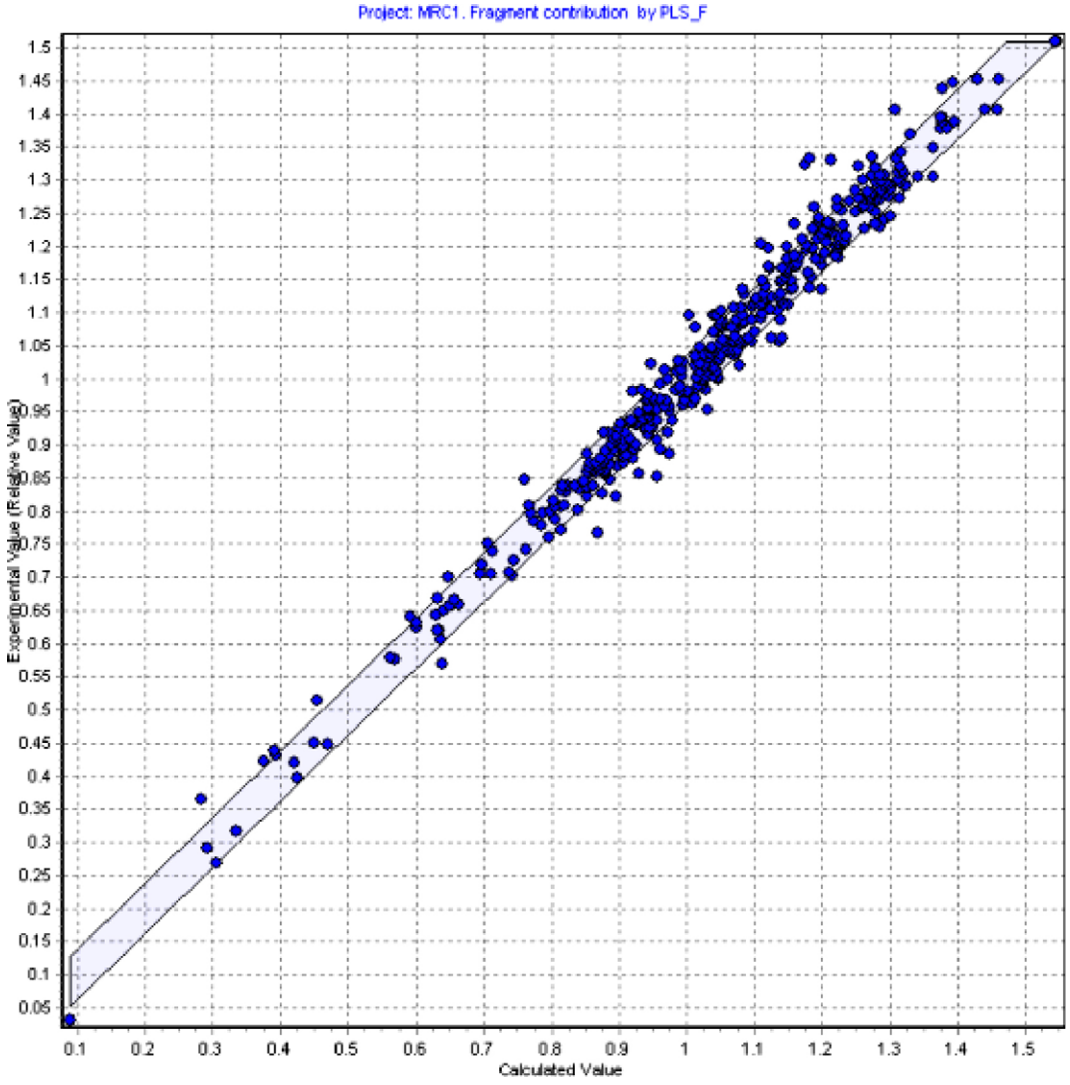

Fig. 8. Experimental/calculated value graph for PLS_F(2) model (blue area is the 95% confidence region).

Table 3
Statistics for *Y*-randomization results

Model	r^2_{min} [a]	r^2_{max} [b]	Chi square$_{min}$ [a]	Chi square$_{max}$ [b]	S.E.$_{min}$ [a]	S.E.$_{max}$ [b]
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
LSF_F(2)	0.5407	0.7146	5.599	9.011	0.193	0.245
PLS_F(2)	0.9307	0.9492	1.023	1.429	0.051	0.060

[a] Minimum value from the 100 iterations.
[b] Maximum value from the 100 iterations.

Table 4
Experimental and calculated values for the external test set

Compound numbers	Experimental value	LSF_F(1)	PLS_F(1)	LSF_F(2)	PLS_F(2)
1	0.940701176	–1.053	0.3593	0.8943	0.3431
2	1.369707832	1.266	0.9104	3.543	1.28
3	1.391028658	1.76	0.8943	4.095	1.205
4	1.247037447	1.522	0.8603	3.061	0.9861
5	1.247541674	1.545	1.012	0.5264	0.9422
6	1.231141038	1.248	1.177	2.888	1.48
7	1.246448298	1.036	1.16	2.339	1.387
8	1.19896607	0.625	0.8795	0.9593	0.9537
9	1.34045207	1.68	1.052	5.367	1.222
10	0.175014419	0.961	0.9435	0.2037	0.6853
11	0.072502486	0.753	0.9676	0.1963	0.6814
12	0.303884138	0.552	0.7173	0.9323	0.6943
13	0.246922448	1.147	0.5971	0.7335	0.5345
14	0.113790688	0.9026	0.842	1.156	0.7553
15	0.184663725	0.4974	0.5755	0.5701	0.6852
16	0.405005467	0.6296	0.7273	0.4984	0.7573
17	1.293383839	1.286	1.111	2.811	1.174
18	1.311164463	1.332	1.102	2.987	1.328
19	0.268338812	1.656	0.9799	1.666	0.7814
20	1.17803801	0.9315	0.8373	1.119	0.894
21	1.175187802	1.257	1.121	3.095	1.406
22	1.079421891	0.2324	0.6321	0.7431	0.3984
23	0.126497199	–0.1488	0.5452	0.928	0.4854
24	0.422595016	0.8974	0.7826	0.8437	0.6507
25	0.7490632	0.9153	0.8748	0.6544	0.6142
26	1.273119235	–0.0366	0.9689	2.395	1.23
27	0.06258249	0.8578	0.9714	0.8154	0.7035
28	0.033024192	0.2679	0.1993	0.3215	0.3056
29	0.040444286	0.0528	0.2263	1.135	0.3547
30	0.03178751	0.1398	0.3715	1.089	0.4038
31	0.424967535	–1.332	0.3874	2.513	0.4496
32	0.163751588	–0.9275	0.3114	2.165	0.3461
33	0.350623295	–1.405	0.3152	2.31	0.4097
34	0.070644809	0.5745	0.4144	0.5711	0.3905
35	0.360851135	0.8101	0.7858	–0.1901	0.4329
36	0.174934804	1.249	0.533	0.6898	0.4935
37	1.166196645	–1.426	0.6206	0.5364	0.4564
38	0.961947695	1.079	1.059	0.7638	1.054
39	1.01634579	1.694	0.9871	0.7456	1.174
40	1.046742696	–0.7477	1.079	2.064	1.11
41	0.934512457	1.083	1.043	0.6883	1.117
42	1.021961282	0.2527	0.8793	1.83	0.8352
43	1.001303912	1.035	1.049	1.396	1.005
44	0.863989724	0.7775	0.9993	2.149	0.8083
45	1.088773968	0.8048	1.059	0.8876	1.195
46	1.158330709	0.734	1.023	1.304	0.9724
47	0.998310398	0.6362	1.022	1.337	1.044
48	1.082001408	1.143	1.096	1.162	0.9682
49	0.948264576	1.047	0.9573	0.9653	1.156
50	1.028081002	1.03	1.002	0.5633	1.068
51	1.000619226	0.7104	1.019	0.8221	1.05
52	0.893452484	1.073	1.019	1.129	0.9884
53	1.080817803	2.465	0.9903	2.412	1.217
54	1.094601767	1.248	1.177	2.888	1.48
55	0.99042854	2.044	0.9125	2.685	1.138
56	1.013315122	0.8281	0.9205	1.014	1.059
57	0.892237033	0.1312	0.9765	1.079	1.021
58	0.977382338	0.5141	0.971	1.289	0.784
59	0.742964711	1.714	1.033	2.241	1.103

Table 5
Statistics for external test set validation

Model	r^2	q^2	Chi square	S.E. [a]	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
LSF_F(2)	0.7648	–5.117	4.615	0.175	0.0087	1.084
PLS_F(2)	0.9771	0.598	0.465	0.003	0.0017	0.757

[a] Standard error.

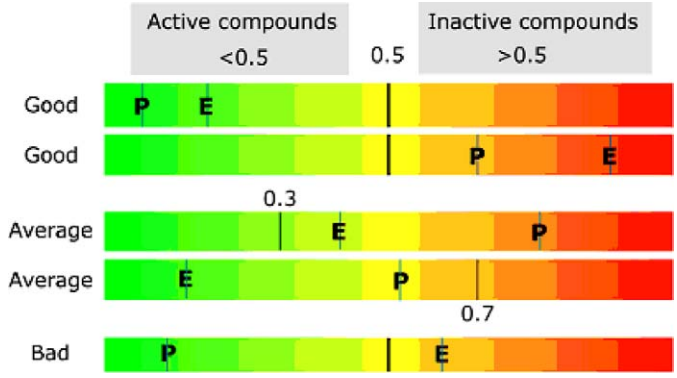

Fig. 9. Classifications of the results from the predictive models (P = predicted activity and E = experimental activity).

Table 6
Classification results

Model	Number of "bad" results	% of "bad" results	Number of "average" results	% of "average" results	Number 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
LSF_F(2)	12	20.3	4	6.8	43	72.9
PLS_F(2)	3	5.1	10	16.9	46	78.0

References

[1] K. Ansell, B. Saxty, C. Kettleborough, M. Dalrymple, J. Corrie, M. Blackman, Poster presentation: Society for Biomolecular Screening Eighth Annual Conference, The Hague, The Netherlands, 22–26 September 2002.

[2] PredictionBase 2.0. ID Business Solutions Ltd., 2 Occam Court, Occam Road, Surrey Research Park, Guildford, Surrey GU2 7QB, http://www.idbs.com/PredictionBase/.

[3] D.M. Hawkins, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.

[4] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, Cambridge MA, USA, 1992 (second ed.).