Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2004) 467–472

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks

M.G. Ford [a,*], W.R. Pitt [b], D.C. Whitley [a]

[a] Centre for Molecular Design, IBBS, University of Portsmouth, King Henry Building, King Henry I St., Portsmouth PO1 2DY, UK
[b] Celltech R&D Ltd., Granta Park, Great Abington, Cambridge CB1 6GS, UK

Accepted 3 March 2004

Abstract

Linear discriminant analysis and a committee of neural networks have been applied to recognise compounds that act at biological targets belonging to a specific gene family, protein kinases. The MDDR database was used to provide compounds targeted against this family and sets of randomly selected molecules. BCUT parameters were employed as input descriptors that encode structural properties and information relevant to ligand–receptor interactions. The technique was applied to purchasing compounds from external suppliers. These compounds achieved hit rates on a par with those achieved using known actives for related targets when tested for the ability to inhibit kinases at a single concentration. This approach is intended as one of a series of filters in the selection of screening candidates, compound purchases and the application of synthetic priorities to combinatorial libraries.

Keywords: Compound selection; Neural networks; High-throughput screening; Discriminant analysis; BCUT; Protein kinases

1. Introduction

The pharmaceutical industry is seeking ways of increasing efficiency, in part through the use of chemical libraries and high-throughput screening. The principal objective is to avoid compounds likely to fail during development, i.e. candidate compounds should fail fast and fail cheap. One approach to improving the efficiency of drug discovery is to predict drug-like characteristics. This is based on the strategy that screening compounds predicted to be drug-like should result in hits/leads with shorter development times and reduced chances of failing. A number of recent studies have attempted to separate potential drugs from non-drugs, using methods ranging from simple physicochemical filters [1,2] to artificial intelligence techniques such as neural networks [3,4], genetic algorithms [5] and decision trees [6].

In spite of these developments a number of problems remain to be addressed. In some cases, compounds are classified as drug-like only if they resemble known classes of drugs, but this can present problems if the intention is to optimise leads in order to avoid existing patents. The aim here is to provide hits with chemical structures that are diverse, thus helping to overcome the issue of patentability, but accessible and plausible to medicinal chemists. Attempts have also been directed at prior prediction of the biological activities of the compounds to be tested during the screening process. One approach is to develop QSAR models to predict the activity of compounds held in large databases, commercial catalogues and virtual libraries. The choice of descriptor can be of crucial importance, and a number of molecular descriptors have been used successfully, including various fingerprints and 3/4-point pharmacaphores. The present work is based on the BCUT metrics developed from the work of Burden and Pearlman [7–9]. These have been used in several drug discovery applications, including ACE inhibitors and the classification of kinase ligands [10]. They encode structural properties and information pertinent to ligand–receptor binding, and were used here for their potential to select compounds with novel structures, a process sometimes called ‘scaffold-hopping’.

The choice of the data analysis procedure is also critical, since the assumptions that underpin a particular methodology must be shown to apply to the data under investigation.

* Corresponding author. Tel.: +44-2392-843-020;
fax: +44-2392-843-722.
E-mail address: martyn.ford@port.ac.uk (M.G. Ford).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2004.03.006

An earlier study [11] described the use of artificial neural networks (ANNs) trained on BCUT descriptors to classify drug-like compounds into subsets that act at biological targets belonging to specific gene families. The purpose of the present work is to present two further results for one of these gene families, protein kinases. First, the performance of the ANNs in terms of a confusion matrix (i.e. the sensitivity, specificity and the predictive power of positive and negative tests) is compared with a linear discriminant analysis. The approach with the highest predictive power, which, not surprisingly, turned out to be the ANNs, was used to select a collection of compounds for purchase. The second aim of the paper is to report experimental results on these compounds.

2. Methodology

Protein kinase ligands were extracted from the MDDR database [12], and a random set of compounds was selected from the same source to provide examples of inactive compounds. To ensure that equal weighting was given to both groups during training a 1:1 ratio of active to inactive compounds was used. A drug-likeness filter [2] was applied to remove unsuitable compounds. The Tanimoto index (TI) was then calculated from Unity fingerprints [13] and compounds with a value greater than 0.8 were removed from the data to leave a more chemically diverse set for training. The Diverse Solutions software [9] was used to calculate 29 standard 3D BCUT metrics for each compound. Redundant variables were removed by applying the unsupervised forward selection (UFS) algorithm [14] with a cut-off value of 0.99. Finally the data was divided randomly into training and test sets in the ratio 4:1, with equal numbers of active and inactive compounds in each subset. The training set was used to derive models, while the test set was used to give an independent indication of the ability of the models to generalise to unseen data.

Linear discriminant analysis (LDA) [15] is a multivariate statistical procedure that aims to split objects into two or more categories. This is appropriate since the intention here is to identify and accept for screening compounds likely to show a particular class of biological activity, but identify and reject compounds unlikely to have this property. The compounds in the training set are assigned y values of 0 or 1 to indicate *inactive* or *active*, respectively, and a discriminant function is constructed in the form of a weighted sum of molecular properties $y = c + \sum b_i x_i$. This is optimised by adjusting c and β_i to obtain a maximum separation of the two classes, with inactive compounds distributed around 0 and active compounds distributed around 1. The aim of LDA is to obtain complete resolution of the objects (compounds). However, discriminant functions usually result in only partial resolution (Fig. 1).

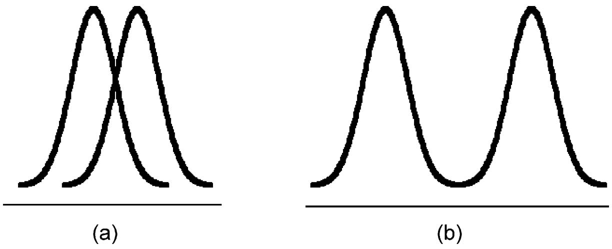

Fig. 1. Partial (a) and complete (b) discrimination of two classes.

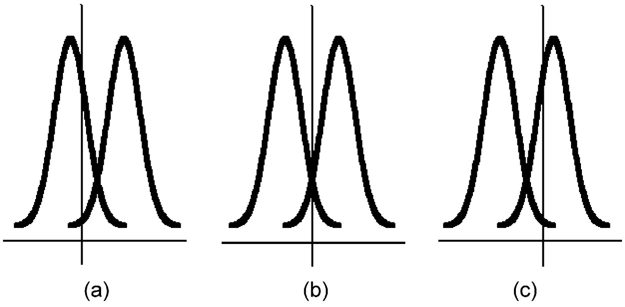

Fig. 2. Choices of cut-off for discriminant function.

The next step is to identify a point M along the discriminant function that provides a boundary. The choice will depend on the objective of the discriminant analysis: should the number of true positives (actives) be maximised, the number of true negatives (inactives) minimised, or some other compromise be adopted? The options are illustrated in Fig. 2. For a given discriminant function and choice of cut-off value M, a confusion matrix can be calculated (Table 1). The entries in this matrix determine four important quantities: sensitivity is the conditional probability that an active is correctly predicted; specificity is the conditional probability that an inactive is correctly predicted; the predictive power of a positive test (PPPT) is the percentage of compounds predicted to be active that are active; and the predictive power of a negative test (PPNT) is the percentage of compounds predicted to be inactive that are inactive. A common procedure is to choose M to maximise the value of $w \times \text{sensitivity} + (1 - w) \times \text{specificity}$, where w is the proportion of active compounds in the training set. In the present context, the appropriate measure appears to be the predictive power of a positive test, since the proportion of

Table 1
Confusion matrix

	Predicted active	Predicted inactive	% Correct	
Active	TP	FN	$\frac{100 \times \text{TP}}{\text{TP} + \text{FN}}$	Sensitivity
Inactive	FP	TN	$\frac{100 \times \text{TN}}{\text{TN} + \text{FP}}$	Specificity
% Correct	$\frac{100 \times \text{TP}}{\text{TP} + \text{FP}}$ PPPT	$\frac{100 \times \text{TN}}{\text{FN} + \text{TN}}$ PPNT	$\frac{100 \times (\text{TP} + \text{TN})}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	

TP: true positive; TN: true negative; FP: false positive; FN: false negative; PPPT: predictive power of a positive test; PPNT: predictive power of a negative test.

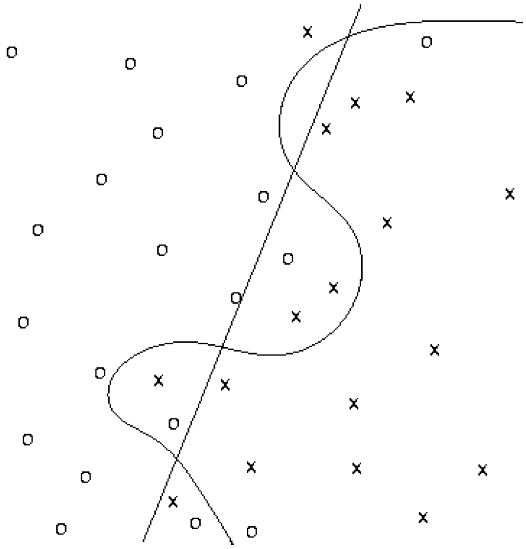

Fig. 3. Linear and non-linear discriminant functions.

compounds predicted to be active that really are active is the principal aim of the screening exercise. In the current study LDA was carried out using Minitab [16].

Although the boundary used in LDA can be adjusted to maximise either sensitivity, specificity, PPPT or PPNT, a linear discriminant function can only achieve complete resolution of active and inactive compounds if the two classes are linearly separable; i.e. the classes are divided by a linear hyper-plane in the descriptor space. In general this is not the case, and a non-linear function is required to improve the separation of the classes (Fig. 3). In this study non-linear discriminant functions were constructed using artificial neural networks.

Feed-forward, back-propagation networks were used with the UFS reduced BCUT descriptors as the input layer, a single layer of hidden units with tanh activation functions, and a single output unit with a logistic activation function. The networks were trained to minimise a cross-entropy error function, with a weight decay regularisation term added to penalise large network weights (and so improve the generality of the network) [17]. Network training was carried out in MATLAB [18] using the NETLAB [19] library. The training protocol required the training set to be further divided into network training and validation sets, in the ratio 3:1, again maintaining equal numbers of active and inactive compounds in each subset. An early-stopping rule based on the validation set was employed to avoid over-training [17]. The training procedure was divided into two stages. First, the optimal network architecture (i.e. the number of hidden units) was determined by training a set of networks with different numbers of hidden units and selecting the architecture which minimised the validation set errors. Second, a set of 1000 networks was trained using the optimal architecture, and the best 100 of these (based on validation set errors) were selected as an ensemble, or committee, of networks to produce the final classification. For details of the training protocol see Manallack et al. [11]. The entire procedure is summarised in Fig. 4.

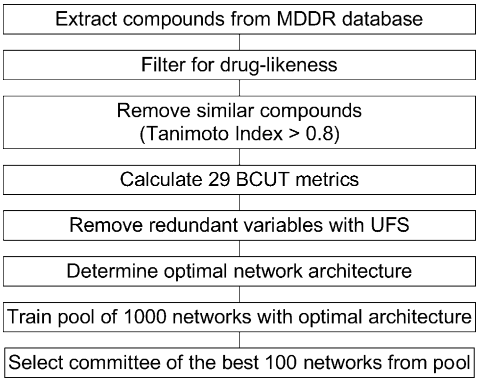
Extract compounds from MDDR database
Filter for drug-likeness
Remove similar compounds
(Tanimoto Index > 0.8)
Calculate 29 BCUT metrics
Remove redundant variables with UFS
Determine optimal network architecture
Train pool of 1000 networks with optimal architecture
Select committee of the best 100 networks from pool

Fig. 4. Network training procedure.

3. Results

Initially 1524 kinase ligands were extracted from the MDDR database but the list was reduced to 240 as a result of the use of drug-likeness filters and the Tanimoto index to increase the chemical diversity of the data. A further 240 compounds were selected at random to represent negative training cases. The resulting 480 compounds were those used in an earlier study [18]. The data were divided into training and test sets containing 384 (80%) and 96 (20%) compounds, respectively.

The LDA confusion matrices for the training and test sets are shown in Table 2. The LDA classified 81.8% of the training set correctly, while this dropped to 75.0% for the test set. Similar falls, between 6 and 8%, were observed between the LDA training and test set performance for each of the measures associated with the confusion matrix: the PPPT, for example, fell from 79.1 to 73.1%.

The subdivision of the training set into network training and validation sets produced sets of 288 and 96 compounds (representing 60 and 20% of the original data, respectively). The first stage of the ANN training regime iden-

Table 2
LDA confusion matrices

	Predicted active	Predicted inactive	Total	% Correct
Training set				
Active	166	26	192	86.5
Inactive	44	148	192	77.1
Total	210	174	384	
% Correct	79.1	85.1		81.8
Test set				
Active	38	10	48	79.2
Inactive	14	34	48	70.8
Total	52	44	96	
% Correct	73.1	77.3		75.0

tified three hidden units as the optimal architecture. A set of 1000 networks with three hidden units was trained and an ensemble of 100 of these with the smallest validation set errors was selected to make the final predictions. The mean of the outputs of these ANNs was taken as the consensus prediction to provide an average discriminant score, on which classification could be based. The confusion matrices based on these average scores for the training and test sets are presented in Table 3. The ANN ensemble classified 83.1% of the training set and 80.2% of the test set correctly. As with the LDA, performance on the test set is weaker than on the training set, as expected, but the difference for the ANNs is less marked: the PPPT falls by just 1%, for example.

As the results are the consensus of 100 networks it is possible to calculate the mean scores and standard deviations (S.D.) for the network outputs. These can then be plotted in order to visualise the degree of certainty associated with the sets of active and inactive compounds and are shown in Fig. 5a for the training set and in Fig. 5b for the test set. By removing compounds with an S.D. greater than 0.1, 0.2 or 0.3 an increase in the confidence of the prediction can be achieved dependant upon the size of the S.D.

Table 3
Network ensemble confusion matrices

	Predicted active	Predicted inactive	Total	% Correct
Training set				
Active	163	29	192	84.9
Inactive	36	156	192	81.2
Total	199	185	384	
% Correct	81.9	84.3		83.1
Test set				
Active	38	10	48	79.2
Inactive	9	39	48	81.3
Total	47	49	96	
% Correct	80.9	79.6		80.2

It is encouraging to note that the distribution of the means and variances are consistent with that of a binomial system, since this is the expected probability density function for a two-state classification. This is confirmed by the observation

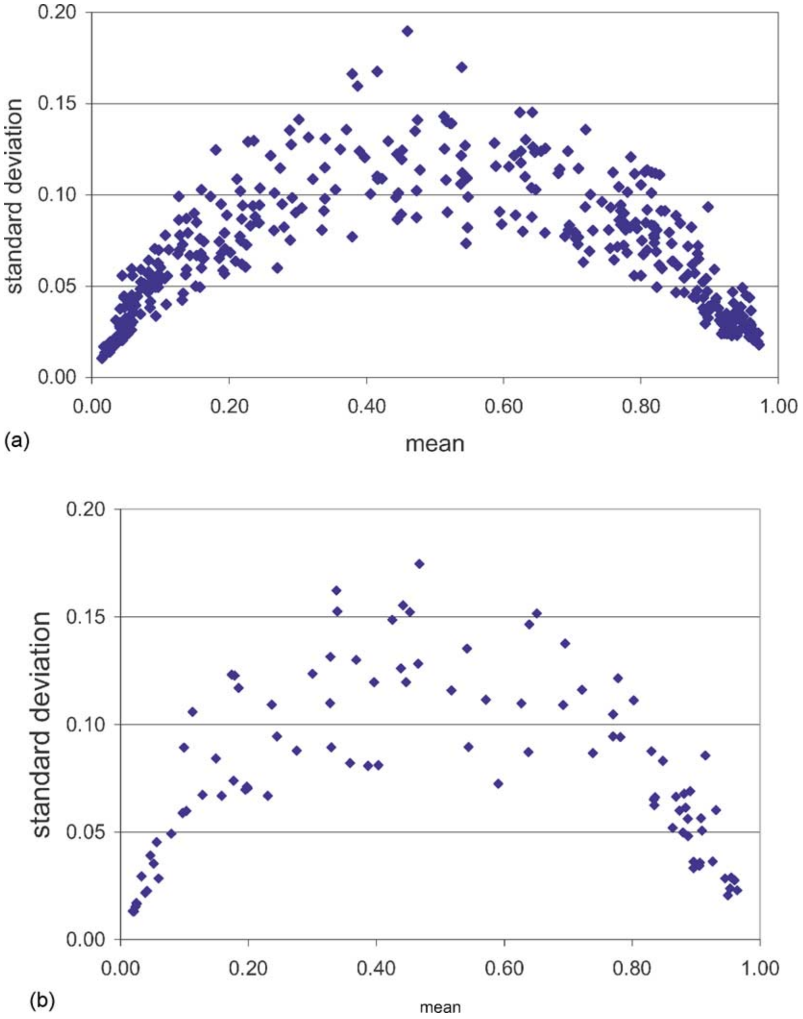

Fig. 5. (a) Distribution of mean scores and standard deviations for the training set. (b) Distribution of mean scores and standard deviations for the test set.

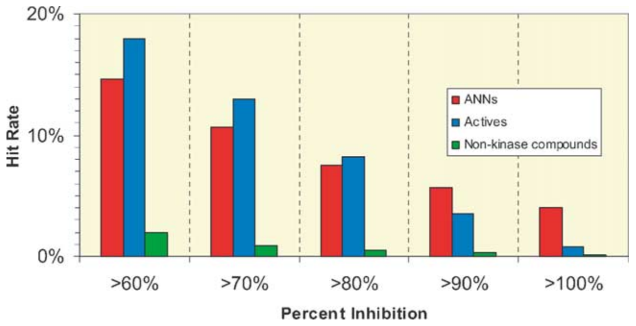

Fig. 6. Experimental kinase inhibition results for focused and non-focused compound sets. Red: compounds purchased after selection by NN; blue: known kinase inhibitors from historical drug discovery programs; green: compounds with no known kinase activity from historical non-kinase drug discovery programs.

that the maximum variance occurs at $p = q = 0.5$ with minima at $p = q = 0$ and $p = q = 1$ (Fig. 5).

Comparing the performance of LDA and the ANNs in Tables 2 and 3 shows that the ANNs have only a marginal (1–2%) advantage on the training set, but that they were more successful on the test set, outperforming LDA by 5% on the total proportion predicted correctly, and by 7.8% on the PPPT. Although based on a relatively small sample size, this improved level of generalisation on the test set led to the adoption of an ANN model for the remainder of the study, and a ‘production’ ANN ensemble was trained, using the entire data set for training.

As a validation of its performance this ensemble was applied to the set of compounds left out at the Tanimoto step, and 99.5% of these compounds were predicted correctly. The BCUT descriptors of these compounds were, of course, very similar to those used to train the ANNs so this result confirms the robustness of the networks for compounds in proximity in the BCUT space to those in the training set.

The ANN ensemble was used to predict the activity of compounds from third-party suppliers, and a library of 816 compounds was purchased from eight different suppliers for screening based on the network results. These compounds were screened along side two sets of in-house compounds. The first of these sets were compounds derived from non-kinase drug discovery programs, which had no known kinase activity (2243 compounds). No quantitative analytical methods were used to select these compounds and they can be considered as a random screening set. The second set contained compounds known to be active against at least one other kinase (602 compounds, labelled actives in Fig. 6). Single point inhibition assays were conducted on these compounds at a fixed concentration (10 μM) in five different protein kinase activity assays. The methodology used for these assays varied but all depended on the detection of the level of phosphorylation of a substrate by changes in the levels of fluorescence of a labelled antibody or substrate. The results of all compounds in all five assays were pooled and are shown in Fig. 6. At a level of 60% inhibition, the hit rate of the set predicted by the ANNs (15%) was almost as high as that for the known actives (18%). By contrast, the non-focused set of compounds achieved a far lower hit rate (2%). Some compounds tested produced an apparent inhibition of greater than 100%. These are normally artificial results, caused by the innate fluorescence of these compounds. Only compounds that gave an inhibition of greater that 60% and were of sufficient interest (novel, chemically tractable, etc.) to the chemists working on the corresponding project were confirmed as active by the measurement of an IC50. Due to the subjectivity of this selection and the limited number of the compounds with a measured IC50, it was felt that single point inhibitions were the best way of gauging the predictive power of the ANN selection method. That said, at the time of writing, 22 ANN selected compounds had been selected for IC50 measurement in at least one of these five assays and were found to have an IC50 < 10 μM.

4. Discussion

ANNs are expected to perform better than LDA, due to the improved separation of the two classes possible with a non-linear discriminant function, and it is perhaps surprising that the two methods showed such similar results on the training set. The ANNs performance on the test set, however, indicates that they are more likely than the LDA to classify new data correctly. The difference is most marked in the results for the predictive power of a positive test, the measure pertinent to compound selection. Indeed the ANNs superior test set results are due to a lower number of false positives; both methods performed identically on the active compounds (numerically, that is—a small number of actives were classified differently by the two methods, but the numbers balance out in the confusion matrices).

The results indicate that pre-processing the data to remove redundancy and reduce multi-colinearity, avoiding non-drug-like compounds and restricting attention to a chemically diverse set of compounds can achieve accurate prediction of protein kinase activity using a committee of ANNs.

Earlier work indicated that committees of ANNs trained on other gene families (amine- and peptide-binding class A rhodopsin-like G protein-coupled receptors) can achieve similar results [11]. The networks’ performance depends on the amount of diversity within the particular gene family: the results for the less diverse protein kinases are rather stronger than those for the more diverse amine family.

The use of committees of ANNs leads naturally to a strategy for in silico screening: predict the activity classes for large numbers of compounds and select those with the most confident predictions (smallest S.D.). From these the chemist can select a further subset that are chemically accessible, have appropriate physical properties and low cost for purchase and subsequent HTS. It is of interest to note that committees of ANNs have also been used to make accurate predictions of a key physical property, water solubility [20].

One difficulty with this approach is the lack of a validated training set, in that the set used as inactive cases during training may, in fact, include some active examples. However, this is likely to increase the number of false negatives, whereas the aim of the investigation is to select compounds predicted to have a higher probability of hitting certain target families, and to develop a probabilistic approach for prioritisation of candidate compounds for synthesis and screening.

Compound collections selected by networks trained to recognise other gene family targets are currently being screened.

5. Conclusions

BCUT metrics have demonstrated utility in discriminating compounds that interact with particular gene families. Consensus neural networks generalise more effectively to unseen data than linear discriminant analysis, and allow confidence levels to be applied to the output results, allowing a probabilistic prioritisation for synthesis and screening.

The trained networks have been used to select compounds to purchase for use in appropriate gene-target screens, and the results obtained from those for the protein kinase family are encouraging.

Acknowledgements

The authors wish to acknowledge the help and advice of David Manallack (DeNovo Pharmaceuticals), Prof. David Livingstone (Chemquest and Portsmouth University), and Dr. David Salt (Portsmouth University) throughout this work, which was funded by Celltech R&D Ltd. The authors would also like to thank the referees for their comments, which led to a substantial improvement of the original paper.

References

[1] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.

[2] M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller, N. Ramsden, Strategic pooling of compounds for high-throughput screening, J. Chem. Inf. Comput. Sci. 39 (1999) 897–902.

[3] Ajay, W.P. Walters, M.A. Murcko, Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J. Med. Chem. 41 (1998) 3314–3324.

[4] J. Sadowski, H. Kubinyi, A scoring scheme for discriminating between drugs and nondrugs, J. Med. Chem. 41 (1998) 3325–3329.

[5] V.J. Gillet, P. Willett, J. Bradshaw, Identification of biological activity profiles using substructural analysis and genetic algorithms, J. Chem. Inf. Comput. Sci. 38 (1998) 165–179.

[6] M. Wagener, V.J. van Geerstein, Potential drugs and nondrugs: prediction and identification of important structural features, J. Chem. Inf. Comput. Sci. 40 (2000) 280–292.

[7] F.R. Burden, Molecular identification number for substructure searches, J. Chem. Inf. Comput. Sci. 29 (1989) 225–227.

[8] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci. 39 (1999) 28–35.

[9] R.S. Pearlman, K.M. Smith, Novel software tools for chemical diversity, Perspect. Drug Discov. Des. 9 (1998) 339–353.

[10] B. Pirard, S.D. Pickett, Classification of kinase inhibitors using BCUT descriptors, J. Chem. Inf. Comput. Sci. 40 (2000) 1431–1440.

[11] D.T. Manallack, W.R. Pitt, E. Gancia, J.G. Montana, D.J. Livingstone, M.G. Ford, D.C. Whitley, Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks, J. Chem. Inf. Comput. Sci. 42 (2002) 1256–1262.

[12] The MDDR database is available from MDL Information Systems Inc., San Leandro, CA, USA.

[13] UNITY, Tripos Inc., 1669 S. Hanley Rd., Suite 303, St. Loius, MO, USA.

[14] D.C. Whitley, M.G. Ford, D.J. Livingstone, Unsupervised forward selection: a method for eliminating redundant variables, J. Chem. Inf. Comput. Sci. 40 (2000) 1160–1168.

[15] W.R. Dillon, M. Goldstein, Multivariate Analysis: Methods and Applications, Wiley, New York, 1984.

[16] Minitab Inc., Quality Plaza, 1829 Pine Hall Road, State College, PA, USA.

[17] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York 1995.

[18] MATLAB, The MathWorks Inc., Natick, MA, USA.

[19] I.T. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, London, Berlin, Heidelberg 2002.

[20] D.T. Manallack, B.G. Tehan, E. Garcia, B.D. Hudson, M.G. Ford, D.J. Livingstone, D.C. Whitley, W.R. Pitt, A consensus neural network-based technique for discriminating soluble and poorly soluble compounds, J. Chem. Inf. Comput. Sci. 43 (2003) 674–679.