Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks

David T. Manallack,*,† Will R. Pitt,† Emanuela Gancia,† John G. Montana,‡ David J. Livingstone,§,# Martyn G. Ford,§ and David C. Whitley§

Celltech R&D Ltd., Granta Park, Great Abington, Cambridge, CB1 6GS, UK, Amedis Pharmaceuticals, Unit 209, Cambridge Science Park, Milton Road, Cambridge CB4 0GZ, UK, Centre for Molecular Design, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, Hampshire, PO1 2DY, UK, and ChemQuest, Delamere House, 1 Royal Crescent, Sandown, Isle of Wight, PO36 8LZ UK

Received March 27, 2002

A series of neural networks has been trained, using consensus methods, to recognize compounds that act at biological targets belonging to specific gene families. The MDDR database was used to provide compounds targeted against gene families and sets of randomly selected molecules. BCUT parameters were employed as input descriptors that encode structural properties and information relevant to ligand−receptor interactions. In each case, the networks identified over 80% of the compounds targeting a gene family. The technique was applied to purchasing compounds from external suppliers, and results from screening against one gene family demonstrated impressive abilities to predict the activity of the majority of known hit compounds.

INTRODUCTION

Numerous technologies have been developed in recent years to improve the efficiency of drug discovery (e.g. high-throughput screening (HTS)[1] and combinatorial chemistry[2,3]). Indeed, the process of finding new medicines has undergone dramatic changes that can be attributed to increasing economic pressures on the pharmaceutical industry.[4] The focus is now oriented toward avoiding compounds that are likely to fail in the expensive final phases of drug development, the so-called 'fail fast, fail cheap' strategy.

The major reasons for the failure of compounds in development are those of inadequate efficacy (≈31%), poor biopharmaceutical properties (≈41%) and toxicity (≈22%).[5,6] To address this, various in vitro assays for absorption, distribution, metabolism, excretion and toxicity (ADMET) are now applied early in the research phase to eliminate unsuitable compounds.[7,8] In addition, much research has been directed at improving *in silico* ADMET predictions[9−11] to guide the selection of screening candidates, purchases from third-party compound suppliers and combinatorial chemistry library design. The obvious advantage of such computational methods is that physical samples of the compounds are not required.

One aspect of improving these *in silico* screening methods has been applied to predicting whether a compound is 'drug-like'. Research on predicting drug-like characteristics includes the application of simple filters using physicochemical properties,[12−15] the use of decision trees[16] and the application of more complex analyses using artificial intelligence.[17−19] Work has also concentrated on defining 'chemical space' to find regions occupied by drug-like compounds[20,21] and analyses have been conducted on the frameworks represented by drugs.[22,23] Two novel scoring schemes, multilevel chemical compatibility (MLCC)[24] and drug-like index (DLI)[25] have been described which can be used as filters on sets of compounds to select those considered to be drug-like. The PASS software[26] was originally developed to predict the biological activity of a compound and has now been adapted for identifying drugs and nondrugs. Despite the outstanding results of attempts to classify drugs and nondrugs, some criticism has been directed at the lack of validated databases used in these studies. For some of these techniques, compounds may only be selected if they resemble existing drugs, making it potentially difficult to avoid patent problems if these compounds are used as starting points for optimization. In an attempt to circumvent this, Muegge and co-workers[27] developed a simple pharmacophore point filter capable of discriminating drugs from nondrugs based on the observation that nondrugs are often under functionalized. Their results compared well to, and were complementary to, neural network methods. Of course, the application of all these methods is intended to result in lists of screening candidates with the potential to afford shorter optimization timelines if found to be hits. Further thought has also been given to the ease of chemical modification of hits emerging from biological screens in addition to patentability.

Studies related to those listed above include attempts to use computational methods to predict the biological activity profiles of compounds, and considerable amounts of work have been oriented toward this. However, we shall focus on more recent studies aimed at predicting profiles of sets of compounds (e.g. commercial supplier's databases) or virtual libraries prior to synthetic work. Among the methods used have been topological patterns[28] and similarity searching[29,30] based on the chemical structure and other complex repre-

* Corresponding author phone: +44 1223 238000; fax: +44 1223 238088; e-mail: David.Manallack@denovopharma.com. Present address: De Novo Pharmaceuticals Ltd. Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, UK.
† Celltech R&D Ltd.
‡ Amedis Pharmaceuticals Unit 209.
§ University of Portsmouth.
# ChemQuest.

10.1021/ci020267c CCC: $22.00
Published on Web 08/09/2002

sentations of molecules, often called fingerprints. Xue and co-workers introduced the concept of short binary bit strings, or minifingerprints, to search for compounds with similar biological activities,[31] suggesting that highly complex fingerprints may be too sensitive to structural variations in compounds with similar biological activity. Another study using 'affinity' fingerprints,[32] not directly derived from molecular structures, was able to classify compounds belonging to five activity classes regardless of structural type. More extensive profiles can be generated on sets of compounds using three and four dimensional pharmacophore descriptors.[33−35] Like fingerprints, these methods use a binning method to indicate the presence or absence of particular three or four point pharmacophores using a set of predefined distances and functional groups. These methods can generate profiles of many millions of datapoints that can be used for molecular comparisons.

Artificial intelligence methods have been used widely to select compounds on the basis of their biological activities. Many of these have used neural networks trained with a variety of descriptors.[36−39] The neural networks trained by Ajay and co-workers[36] to recognize CNS-active compounds were subsequently applied to the design of a small library of potential CNS-active compounds suitable for synthetic work. Genetic algorithms have also been used,[40,41] in addition to binary kernel discrimination,[42] a method that deals particularly well with noisy data. Harper et al.[42] also pointed out that their system is capable of dealing with very structurally diverse compounds acting through different mechanisms which gets around the problem of selecting compounds that resemble known actives. Their technique, albeit a 'black box', is designed not to suggest structural changes but to provide an estimate of the chances that a molecule is active. Indeed, the majority of these methods are not aimed at finding a development candidate but are intended to select compounds that have been predicted to have a high probability of hitting the target being studied. They are expected to complement other methods such as pharmacophore generation and searching, in addition to structure-based design work.

One particularly interesting set of descriptors we have explored in this current study are BCUT metrics. These descriptors, stemming from the work of Burden[43] and Pearlman and Smith,[20,21] have been implemented in the DiverseSolutions (DVS) software.[44] From the set of BCUT descriptors generated, the software can be used to select automatically the subset of BCUTs that best represent the structural diversity of the data set being studied. This has been used to highlight regions of space (receptor relevant sub-spaces[20]) in which active compounds cluster. Further applications of the technique have emerged[45−49] indicating that these descriptors appear to be extremely useful for drug discovery. Pirard and Pickett[50] were able to use BCUTs to classify kinase inhibitors active against five different protein kinases. This study was particularly interesting as it suggested that the BCUTs were not merely encoding 2D structural information but contained information relevant to ligand−receptor interactions. Moreover, our interest in these parameters was strengthened as we sought descriptors that have the potential to select compounds that fall outside of known structural classes for a particular target or set of targets, a technique that has been called 'scaffold-hopping'.[51]

The work we shall describe here covers the novel application of neural networks (NNs) as a screening selection tool to identify compounds that target specific gene families. We have considered three target families and show how BCUTS can be used as input parameters for NN training. Furthermore these networks were applied to the selection of screening candidates suitable for in-house assay systems in parallel with filters to remove non drug-like compounds. Currently, the pharmaceutical industry is considering more and more targets resulting from information from the human genome project, many of which belong within gene families. The advantage of our approach is that screening libraries of compounds can be accumulated as a resource for biological targets that belong to the same gene family.

METHODS

Compound Selections. Compounds were extracted from the MDDR database[52] based on the relevant internal 'activity' code field for the gene family of interest. (e.g. 78373, which corresponds to a protein kinase C inhibitor). The term 'gene family' was applied to sets of targets whose genetic origins can be linked together (i.e. through the use of phylogenetic trees). In the case of GPCR's we considered Class A Rhodopsin like receptors, subcategories *amine* and *peptide*.[53] In some cases within the MDDR database, compounds were simply classified under broad activity fields such as 'analgesic' or 'anti-cancer'. This was particularly problematic for gathering compounds acting at kinase targets. In this case, a number of reviews were used to identify scaffolds known to act at kinase enzymes to allow substructure searching.[54,55] Care was taken to ensure that the compounds extracted were indeed kinase inhibitors and not acting through an alternative mechanism. Similarly, chemokine ligands[56] were extracted using substructure searches. In all, three families of compounds were obtained for further study; protein kinase inhibitors, compounds acting at class A rhodopsin-like amine GPCR's (e.g. adrenoceptors, dopamine, 5HT, serotonin, etc.), and compounds acting at class A rhodopsin-like peptide-binding GPCR's (e.g. chemokine, CCK, opioid, etc.).

These compounds would be used to train NNs to recognize compounds with a particular biological profile, and the resulting NNs applied to select screening candidates from third-party compound suppliers and in-house collections, and to direct combinatorial synthetic work. Therefore, filters were applied to remove non drug-like molecules. Simple filters were used such as those described by Hann *et al.*[12] The MDDR database often contains scaffold structures for which there may be many analogues (e.g. compounds from patents) and we felt it appropriate to reduce the data set in size by the application of a simple Tanimoto cutoff of 0.8 (based on Unity fingerprints). Typically this reduced the data sets in size by 85%. The intention of this step was to avoid the networks seeking simple solutions by focusing on sets of compounds highly similar to each other. Following this Tanimoto step the data set used for final training was more diverse thus adding a level of challenge to the network training. A final inspection of each compound list was undertaken by an experienced medicinal chemist to remove any compounds deemed unsuitable for training purposes. The NN training method required compounds not belonging to the gene family as negative cases. A random number

generator was used to select appropriate numbers of compounds from the MDDR database. Compounds in the random set that were common to the gene family list were discarded, and simple drug-like filters[12] were applied to the remainder.

Using the DVS software[44] as interfaced within the Tripos software, 29 standard BCUT descriptors (standard 3D H-suppressed BCUTs) were calculated for each compound. Compounds with missing values were discarded and equal numbers of compounds in the gene family set (positives) and random set (negatives) were used for subsequent NN training (i.e. a 1:1 ratio).

Variable Selection. The unsupervised forward selection (UFS) procedure[57] was applied to reduce the level of multicollinearity in each data set. UFS is a forward-stepping algorithm that selects a subset of variables with a minimal amount of multicollinearity. The first two variables are selected to have the smallest pairwise correlation coefficient, the third variable is selected to have the smallest squared multiple correlation coefficient with the first two, and so on. The procedure halts when each remaining unselected variable has a squared multiple correlation coefficient above some preset cutoff value with those already chosen. The algorithm was applied with a cutoff of 0.99 so that only redundant variables and those with a high degree of multicollinearity were removed. This reduced the number of BCUT variables from 29 to 20 for the kinase data and to 21 for both the amine and peptide data sets.

Neural Network Methods. The BCUT data was used to train feed-forward, back-propagation networks with a single layer of hidden units and a single output unit. A logistic activation function was used for the output unit, with targets of 0 and 1 for the negative and positive cases, respectively, while the hidden units had *tanh* activation functions. The networks were trained by the scaled conjugate gradient method to minimize the cross-entropy error function with a weight decay regularizer.[58] The calculations were performed in MATLAB[59] using the NETLAB software library.[60]

Preliminary experiments indicated that similar results were obtained for a wide range of values of the weight decay parameter, α, and this value was set at $\alpha = 0.1$ for all further calculations.

The training procedure contained three stages. First, the selected compounds were divided randomly into training, validation and test sets in the ratio 50%:25%:25%, preserving the 1:1 ratio between positive and negative cases in each subset. Second, the optimal network architecture (the number of units in the hidden layer) was determined by allowing the size of the hidden layer to vary from 2 to 20 hidden units, training 20 networks (with randomly chosen starting weights) using the training set, and selecting the architecture with the best performance on the validation set. Two measures were employed to assess the performance of the networks on the validation set: the mean square error and the area under the receiver operating characteristic (ROC) curve. In fact these measures produced very similar results on the data studied here: in each case three hidden units gave the optimum validation set performance. The test set was not involved in the training process and used only to give an independent estimate of the performance of networks with the optimal architecture. In the third stage of the training procedure 1000 networks with the optimal architecture were trained, using all the data for training, and from these an ensemble, or committee, of 100 networks with the smallest training set error was selected. The mean of the outputs from the ensemble networks was used to make predictions on unlabeled data, while the variance of the outputs provides a measure of confidence in the predictions.

Confusion Matrices. The networks have a single output unit whose value is the probability that a set of input variables represents a positive compound: those with output values greater than 0.5 are classified as positive and those with outputs less than 0.5 are classified as negative. The networks are trained to maximize the total number of compounds (both active and inactive) in the training set that are classified correctly.

To show how the networks perform on the positive and negative compounds separately, the confusion matrices for the best networks are shown in the form:

TP	FN	TP+FN	100*TP/(TP+FN)
FP	TN	FP+TN	100*TN/(TN+FP)
TP+FP	FN+TN	TP+TN+FP+FN	
100*TP/(TP+FP)	100*TN/(FN+TN)		100*(TP+TN)/(TP+TN+FP+FN)

Here TP = true positive, FP = false positive, TN = true negative, and FN = false negative. (For positive/negative read active/inactive.) The entries in the last column are the *sensitivity* (the percentage of positives predicted correctly) and the *specificity* (the percentage of negatives predicted correctly). The entries in the last row are the *predictive power of a positive test* (the percentage of cases correctly predicted to be positive) and the *predictive power of a negative test* (the percentage of cases correctly predicted to be negative). The entry at the bottom right is the total percentage predicted correct. This is the quantity maximized by the network training using an output value of 0.5 to determine the classification.

RESULTS

Kinase Data Set. One thousand five hundred and twenty-four compounds were extracted from the MDDR database for the kinase data set. Following the removal of duplicates, filtration of compounds with undesirable physicochemical properties and functional groups, and the application of a 0.8 Tanimoto, we were left with 242 compounds. BCUTs could not be calculated for two compounds leaving a set of 240 for training purposes. A random set of 240 compounds was selected and processed to represent the negative cases. The results for the second stage of network training are shown in Figure 1. As can been seen from the plot, the error in the validation set rises as the number of hidden layer units is increased, tailing off after about seven hidden units. This behavior was common to the three cases we studied. For our purposes we sought to select the number of hidden layer units that gave the lowest error which corresponded to two units in this case. However, when a set of 1000 2-hidden unit networks was trained, it emerged that many of the networks had the same training set error, indicating that the network training had converged to a small number of local minima. To circumvent this problem we chose to use 3-hidden-unit networks for the kinase set. The confusion matrices for the optimal 3-hidden-unit network used in the second stage of training are given in Table 1. For the 25% of compounds left out as a test set, the network was able to predict 79% of these compounds correctly. This encouraging

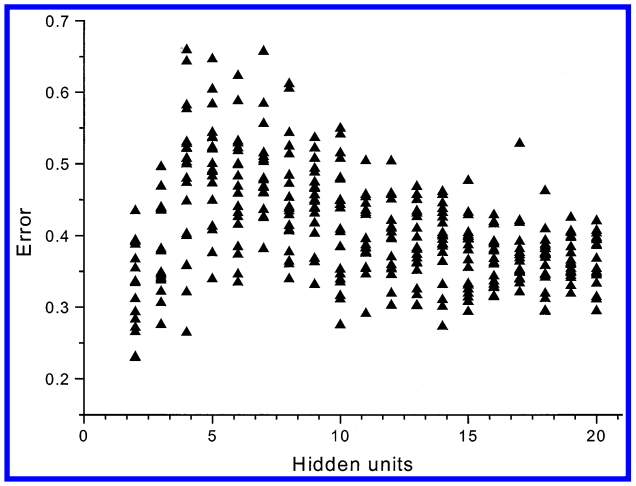

Figure 1. Validation set error vs number of hidden units for initial kinase set training.

Table 1. Confusion Matrices for the Optimal 3-Hidden-Unit Network Using the Kinase BCUT Data

	active	inactive	total	% correct
		Training Set Predictions		
active	105	15	120	87.50
inactive	5	115	120	95.83
total	110	130	240	
% correct	95.45	88.46		91.67
		Validation Set Predictions		
active	45	15	60	75.00
inactive	4	56	60	93.33
total	49	71	120	
% correct	91.84	78.87		84.17
		Test Set Predictions		
active	44	16	60	73.33
inactive	9	51	60	85.00
total	53	67	120	
% correct	83.02	76.12		79.17

Table 2. Confusion Matrix for the Ensemble of 100 Networks Using the Kinase BCUT Data

	active	inactive	total	% correct
		Kinase Consensus Networks		
active	225	15	240	93.75
inactive	12	228	240	95.00
total	237	243	480	
% correct	94.94	93.83		94.38

result led us to complete the third stage of network training using 3 hidden units. From the 1000 3-hidden-unit networks the best 100 were selected as the ensemble for use in screening set selections. The confusion matrix for the ensemble is given in Table 2, which shows that over 94% of compounds were classified correctly.

As the results are the consensus of 100 networks it is possible to calculate the mean output result for each compound and the standard deviation of these 100 results. Figures 2 and 3 show the distribution of the mean and standard deviations for the 480 compounds used in kinase training (240 positives and 240 negatives). The distribution of the means is clearly bimodal, with the majority of results close to either 1 or 0 (Figure 2). The distribution of standard deviation values increased toward zero (Figure 3). This result suggested that some compounds were consistently predicted positive or negative and a smaller population with higher standard deviations could not be predicted consistently across

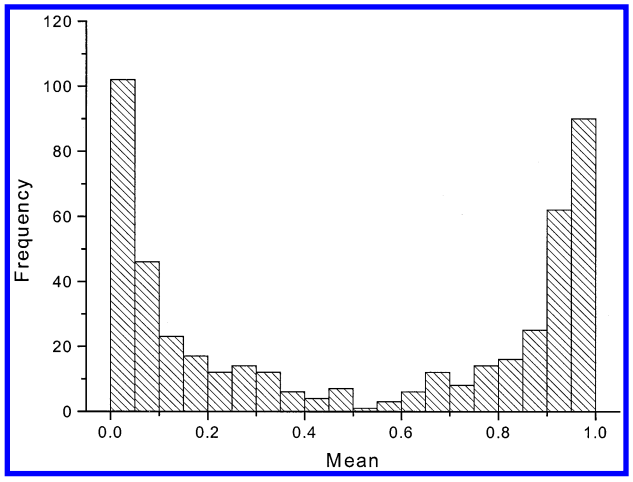

Figure 2. Distribution of mean output values for the 100 kinase consensus networks.

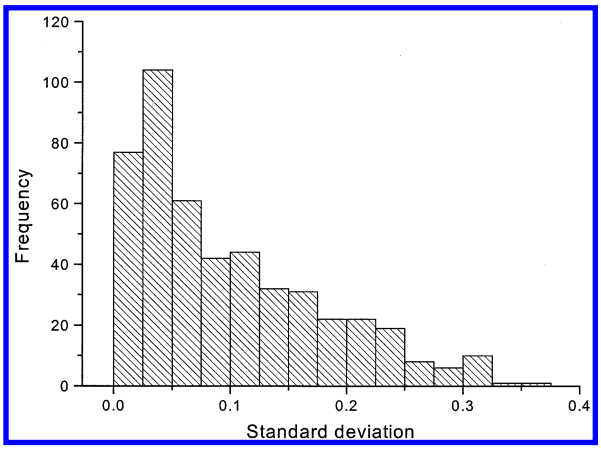

Figure 3. Distribution of standard deviation output values for the 100 kinase consensus networks.

the 100 networks. To test this hypothesis, three series of compounds were selected with standard deviations less than 0.1, 0.2 and 0.3 and their mean values plotted (Figure 4). This reduced the data set to 284, 413 and 468 compounds, respectively. As the standard deviation is reduced it is clear that distribution of the means is separated into 2 groups with no compounds showing a mean value close to 0.5. It follows that these compounds are consistently being classified as positive and negative and for the 284 compounds (0.1 standard deviation) the networks predicted all but one compound correctly. Using a 0.2 cutoff for the standard deviation only six compounds were misclassified.

As part of our validation procedures we first predicted the activity of the compounds left out of the training procedure at the Tanimoto 0.8 step. In this case 916 of the 920 compounds were predicted to be kinase ligands. While this appears impressive it should be remembered that these compounds are all similar to compounds used in the training procedure. An additional data set was also predicted for its kinase profile which comprised a set of 2267 compounds screened in-house against a panel of kinase enzymes. Of these compounds which were either synthesized or purchased specifically for kinase screening, almost all (98.6%) were predicted to be kinase ligands. Approximately one-third of this collection have an IC_{50} value below 1 μM on at least one of the panel of kinase enzymes screened. As a check regarding the similarity of the compounds in the NN training

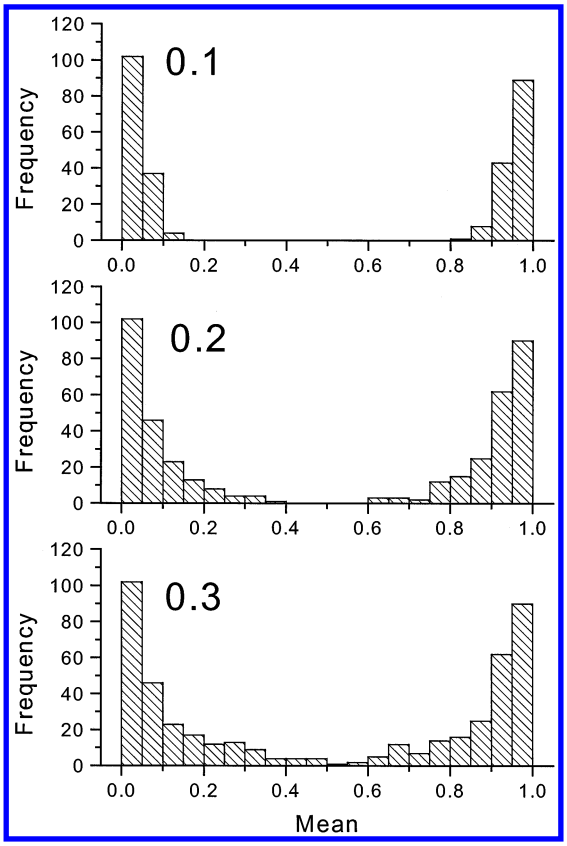

Figure 4. Distribution of mean output values for the 100 kinase consensus networks using cutoffs of 0.1, 0.2 and 0.3 for the standard deviation.

set of 240 kinase ligands and the in-house screening set of 2267 compounds, a comparison was made using the UNITY software.[61] At a similarity of 0.57 and above, 50% of the compounds in the NN training set were similar to one (or more) compounds in the screening set. This was an interesting result as it suggests that 50% of the compounds in the screening set are significantly different in structure to the NN training compounds supporting the argument that BCUT metrics are useful for scaffold-hopping. Indeed, 20% of this 'dissimilar' set of compounds were active ($IC_{50} < 1\ \mu M$) against at least one of the kinase enzymes screened compared to one-third for the entire set of compounds (vide supra).

Two further kinase sets of compounds were predicted using the consensus networks. Both of these came from the application of the PASS software[26] on the NCI database[62] selecting protein kinase C and tyrosine kinase inhibitors. For protein kinase C only 16 compounds were available for study and the consensus networks predicted 12 of these to be kinase ligands, while all of the 281 tyrosine kinase inhibitors were predicted correctly. A number of other data sets were also put through the networks obtained from commercial suppliers as well as the NCI database.[62] In these cases the percentage of compounds predicted to be kinase ligands ranged between 13 and 17%. This number, however, was reduced to approximately 1% or less when a standard deviation cutoff of 0.1 was applied to the output. Of course it cannot be known that these compounds are kinase active without screening them.

GPCR Amine Data Set. A total of 9709 compounds were extracted from the MDDR database and used as the GPCR amine data set. Following the removal of duplicates, compounds with undesirable physicochemical properties and functional groups, and the application of a 0.8 Tanimoto, 1369 compounds remained. The final data set contained 1368 compounds as the BCUT metrics for one compound could not be calculated. A random set of 1368 compounds was used as the negative set. During stage two of network training, networks with 3 hidden units produced optimal results on the validation set, and the test set results for the best 3-hidden-unit network classified 76% of compounds correctly. The confusion matrix for the 100 consensus networks is shown in Table 3 showing that over 80% of compounds were classified correctly. Testing of the compounds left out of the analysis at the Tanimoto step found that 5828 of 7958 (73.2%) were predicted to be ligands of amine GPCRs, a result close to that for the consensus networks (81.5%) shown in Table 3.

Table 3. Confusion Matrix for the Ensemble of 100 Networks Using the GPCR Amine BCUT Data

	active	inactive	total	% correct
	GPCR Amine Consensus Networks			
active	1106	262	1368	80.85
inactive	245	1123	1368	82.09
total	1351	1385	2736	
% correct	81.87	81.08		81.47

Table 4. Confusion Matrix for the Ensemble of 100 Networks Using the GPCR Peptide BCUT Data

	active	inactive	total	% correct
	GPCR Peptide Consensus Networks			
active	609	91	700	87.00
inactive	75	625	700	89.29
total	684	716	1400	
% correct	89.04	87.29		88.14

GPCR Peptide Data Set. The final GPCR peptide data set used in NN training was taken from a set of 9414 compounds extracted from the MDDR database. Following the filtration and Tanimoto steps, 700 compounds were used to calculate BCUT descriptors. An equal number of randomly selected compounds were used as the negative cases. The second stage of network training once again showed that networks with 3 hidden units had the best performance on the validation set, and the test set results for the best 3-hidden-unit network classified 81% of compounds correctly. Table 4 shows the confusion matrix for the 100 consensus networks. Once again the compounds left out at the Tanimoto step were passed through the consensus networks, predicting 1702 of 1955 (87.06%) correctly.

DISCUSSION AND CONCLUSIONS

In this study we have investigated the use of consensus neural network methods and BCUT descriptors to distinguish compounds targeted toward specific gene families. In each of the three cases studied, reasonable predictions could be made relative to randomly selected compounds. These consensus networks have been applied to the purchasing of compounds from third party suppliers for use in biological assays. Further drug-like filters[12] were applied to the purchasing sets as well as similarity checks against compounds already present in our collection. A final check prior to purchasing was undertaken to inspect each compound with an experienced medicinal chemist to weed out further

unwanted structures. One of the attractive aspects of using NNs with BCUT descriptors was the possibility of selecting screening candidates not structurally related to known gene family ligands.

The targets studied here have been termed tractable targets as their function can be interfered with by small molecules and there are a considerable number of examples where drugs have emerged from these gene families. All three can be considered important gene families for therapeutic intervention. The technique we have described, however, could be applied to other gene families such as proteases or ion channels and may be useful for predicting physicochemical properties. To use the method there is a need for a reasonably sized data set for training purposes. The smallest set studied here was 240 compounds, and further work would be needed to determine a recommended minimum number of compounds.

Our work has aimed to investigate a number of issues associated with artificial intelligence and the prediction of biological activities. The first of these arose from the two key studies using neural networks to distinguish drugs and nondrugs.[17,18] In both these cases and related studies, some criticism has been directed toward the lack of validated data sets. In other words, there is the possibility that drugs may be lurking in the databases used as the negative cases (e.g. Available Chemicals Directory[52]). Of course a similar criticism could be applied to this present study, and there is little defense against this unless the entire database is screened against a battery of relevant targets. However, given that the final use of the networks is to select compounds predicted to have activity at particular targets and not to produce a model with 100% accuracy, we are not too concerned about this problem. If this work is viewed as producing probabilistic models then these data sets suffice.

Our second concern was over the use of single neural network models for predictive work. From our observations and work in the neural network field, training the same data set using different starting weights frequently leads to different 'local minima'. As such, predictions made using a single network may not be representative of a population of networks trained on the same set. To address this a consensus (or committee) of networks is used to get the mean and standard deviation of output values. This is a particularly useful method as it is possible to determine which cases are consistently predicted to be one class or the other. More confidence can be applied to these compounds, and this can also be applied during the selection of compounds for purchasing. Committees of networks are discussed in more detail by Bishop,[58] who indicates how the performance of a committee may exceed that of the best single network trained on a particular data set. A third issue, discussed above, is the use of descriptors that allow scaffold-hopping, such as BCUTS. For further discussion about the utility of BCUT metrics, the reader is referred to the recent papers by Pearlman and Smith.[20,21]

The results presented here demonstrate that the neural networks are providing reasonable models predicting over 80% of compounds correctly. The kinase data set performed particularly well on the validation data sets, and this may be a consequence of the limited number of scaffolds that kinase ligands exhibit.[54,55] In contrast to this, GPCR ligands have been studied for many decades, and numerous structural types have been explored. This may account for the reduced performance for the GPCR amine data set (Table 3). One concern we had with the results of the kinase networks was the high percentage of compounds predicted to be kinase active in both the validation sets and commercial suppliers catalogues. For the latter this was around 15%. Lowering the standard deviation to 0.1 reduced this to less than 1% providing more tractable numbers of compounds to deal with for purchasing purposes, with the added advantage that these compounds were consistently predicted to be active.

In this study, the neural networks have acted as a statistical engine to best utilize those BCUT metrics with biologically relevant information. Unfortunately, it is not a simple matter to determine which metrics are involved, and this may be seen as a criticism of the method leading it to be termed a 'black box'. For our purposes though, this was not an important issue as we only sought to select compounds for screening predicted to have a higher probability of hitting targets from the gene family of interest, rather than needing the information to direct lead optimization/design. Future work may, however, seek to investigate which BCUT metrics are involved.

In summary, we have described the novel use of BCUT descriptors in conjunction with consensus neural network methods to successfully discriminate compounds belonging to particular gene families. While previous studies have investigated the prediction of compound activities for specific biological targets, our study has successfully taken on the more challenging task of exploring active molecules within gene families. This work demonstrates that the BCUT metrics contain information relevant to interactions with biological targets in accord with previous studies.[20,21,45−49] In all three cases studied, compounds were selected for purchasing from third party suppliers for use in biological screening. This method is relatively quick to put into operation and is complemented by the benefits of using consensus methods and the use of metrics that allow scaffold-hopping.

ACKNOWLEDGMENT

The authors would like to thank Drs. David Salt and Brian Hudson for their valuable discussions and suggestions.

REFERENCES AND NOTES

(1) Wölcke, J.; Ullmann, D. Miniaturized HTS Technologies − uHTS. *Drug Discov. Today* **2001**, *6*, 637−646.
(2) Furka, A. Combinatorial Chemistry: 20 Years On. *Drug Discov. Today* **2002**, *7*, 1−4.
(3) Kassel, D. B. Combinatorial Chemistry and Mass Spectrometry in the 21st Century Drug Discovery Laboratory. *Chem. Rev.* **2001**, *101*, 255−267.
(4) Drews, J. Jürgen Drews Discusses the Future of the Industry. *Drug Discov. Today* **2001**, *6*, 338−341.
(5) Lipper, R. A. How Can We Optimize Selection of Drug Development Candidates From Many Compounds at the Discovery Stage? *Modern Drug Discovery* **1999**, *2*, 55−60.
(6) Venkatesh, S.; Lipper, R. A. Role of the Development Scientist in Compound Lead Selection and Optimization. *J. Pharm. Sci.* **2000**, *89*, 145−154.
(7) Li, A. P. Screening For Human ADME/Tox Drug Properties in Drug Discovery. *Drug Discov. Today* **2001**, *6*, 357−366.
(8) Thompson, T. N. Early ADME in Support of Drug Discovery: The Role of Metabolic Stability Studies. *Curr. Drug Metab.* **2000**, *1*, 215−241.
(9) Clark, D. E. Prediction of Intestinal Absorption and Blood-Brain Barrier Penetration by Computational Methods. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 477−496.

(10) Johnson, D. E.; Wolfgang, G. H. I. Predicting Human Safety: Screening and Computational Approaches. *Drug Discov. Today* **2000**, *5*, 445−454.
(11) Stewart, B. H.; Wang, Y.; Surendran N. In *Annual Reports in Medicinal Chemistry Volume 35*; Doherty A. M., Ed.; Academic Press: San Diego, 2000; Vol. 35, Chapter 27, pp 299−307.
(12) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden N. Strategic Pooling of Compounds For High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(13) Rishton, G. M. Reactive Compounds and in vitro False Positives in HTS. *Drug Discov. Today* **1997**, *2*, 382−384.
(14) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An Overview. *Drug Discov. Today* **1998**, *3*, 160−178.
(15) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(16) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(17) Ajay, Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish Between "Drug-Like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(18) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(19) Frimurer, T. M.; Bywater, R.; Nærum, L.; Nørskov Lauritsen, L.; Brunak, S. Improving the Odds in Discriminating "Drug-Like" From "Non Drug-Like" Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(20) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(21) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Persp. Drug Discov. Design* **1998**, *9*, 339−353.
(22) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(23) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(24) Wang, J.; Ramnarayan, K. Toward Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524−533.
(25) Xu, J.; Stevenson, J. Drug-Like Index: A New Approach to Measure Drug-Like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(26) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating Between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432−2437.
(27) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-Like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(28) Gálvez, J.; de Julián-Ortiz, J. V.; García-Domenech, R. General Topological Patterns of Known Drugs. *J. Mol. Graph. Mod.* **2001**, *20*, 84−94.
(29) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(30) *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(31) Xue, L.; Godden, J. W.; Bajorath, J. Database Searching for Compounds With Similar Biological Activity Using Short Binary Bit String Representations of Molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(32) Briem, H.; Lessel, U. F. In vitro and *in silico* Affinity Fingerprints: Finding Similarities Beyond Structural Classes. *Persp. Drug Discov. Design* **2000**, *20*, 231−244.
(33) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(34) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation II: Evaluating Molecular Diversity. *Drug Discov. Today* **1996**, *1*, 71−78.
(35) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme C.; Labaudiniere, R. F. New 4-point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(36) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries With CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.
(37) Keserû, G. M.; Molnár, L.; Greiner, I. A Neural Network Based Virtual High Throughput Screening Test for the Prediction of CNS Activity. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 535−540.
(38) Bucinski, A.; Nasal, A.; Kaliszan, R. Pharmacological Classification of Drugs Based on Neural Network Processing of Molecular Modelling Data. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 525−533.
(39) Burden, F. R.; Winkler, D. A. New QSAR Methods Applied to Structure−Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236−242.
(40) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(41) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Components Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(42) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernal Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(43) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(44) DiverseSolutions, v4.0.6; University of Texas, Austin, U.S.A.; distributed through Tripos, Inc.: 1669 S. Hanley Rd., Suite 303, St. Louis, MO 63144, U.S.A.
(45) Beno, B. R.; Mason, J. S. The Design of Combinatorial Libraries Using Properties and 3D Pharmacophore Fingerprints. *Drug Discov. Today* **2001**, *6*, 251−258.
(46) Gao, H. Application of BCUT Metrics and Genetic Algorithm in Binary QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402−407.
(47) Mason, J. S.; Beno, B. R. Library Design Using BCUT Chemistry-Space Descriptors and Multiple Four-Point Pharmacophore Fingerprints: Simultaneous Optimization and Structure-Based Diversity. *J. Mol. Graph. Model.* **2000**, *18*, 438−451.
(48) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.
(49) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(50) Pirard, B.; Pickett, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(51) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.
(52) MDDR and ACD databases are available from MDL Information Systems Inc., San Leandro, CA 94577, U.S.A.
(53) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. GPCRDB: an Information System for G Protein-Coupled Receptors. *Nucleic Acids Res.* **1998**, *26*, 275−279 (see also http://www.cmbi.kun.nl/7tm/phylo/phylo.html).
(54) Myers, M. R.; He, W.; Hulme, C. Inhibitors of Tyrosine Kinases Involved in Inflammation and Autoimmune Disease. *Curr. Pharm. Des.* **1997**, *3*, 473−502.
(55) Boschelli, D. H. Small Molecule Inhibitors of Receptor Tyrosine Kinases. *Drugs Fut.* **1999**, *24*, 515−537.
(56) Trivedi, B. K.; Low, J. E.; Carson, K.; LaRosa G. J. *In Annual Reports in Medicinal Chemistry*; Doherty A. M., Ed.; Academic Press: San Diego, 2000; Vol. 35, Chapter 17, p 191.
(57) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: a Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.
(58) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: 1995.
(59) MATLAB is available from The MathWorks, Inc., Natick MA.
(60) Nabney, I. T. *NETLAB: Algorithms for Pattern Recognition*; Springer, 2002.
(61) UNITY − Tripos, Inc.: 1669 S. Hanley Rd., Suite 303, St. Louis, MO 63144, U.S.A.
(62) NCI database searchable from the following url: http://131.188.127.153/services/ncidb2/.

CI020267C