New QSAR Methods Applied to Structure−Activity Mapping and Combinatorial Chemistry

Frank R. Burden*,† and David A. Winkler‡

Department of Chemistry, Monash University, Clayton 3168, Australia, and CSIRO Division of Molecular Science, Private Bag 10, Clayton South MDC, Clayton 3169, Australia

Received April 15, 1998

A comparison is made of a number of computationally efficient molecular indices with a view to the screening of very large virtual data sets of molecules. The use of Bayesian regularized neural networks is discussed, and their virtue in eliminating the need for validation sets, and potentially even test sets, is emphasized. The concept of a virtual receptor is introduced, and this is illustrated by the results of screening a database of 40 000 molecules.

INTRODUCTION

Historically, bioactive molecules have been discovered by one of several methods:[1] exploratory chemical synthesis; screening of natural products; chemical analoguing of lead compounds; observation of side effects; and serendipity. Although these methods have been, and continue to be, very successful, accounting for the large majority of drugs and agrochemicals discovered to date, they are becoming increasingly expensive and time-consuming. As the efficacy of existing bioactive compounds increases, it becomes more difficult to discover new chemical entities with substantial advantages. The average number of compounds synthesized in order to obtain a commercial candidate has risen from 10 000 to around 40−50 000. The recently-developed combinatorial methods greatly increase the numbers of compounds synthesized and tested but generate very large amounts of data. Clearly it has become very important to find new methods for extracting useful molecular design information from these large quantities of structure−activity data.

The data sets which derive from combinatorial chemistry and high throughput screening are often so massive that QSAR is the method of choice. The method, using multivariate statistics, was developed by Hansch and Fujita,[2] and it has been successfully applied to many drug and agrochemical design problems. In its simplest form the biological response is assumed to be a function of a number of molecular parameters which correlate with molecular size, lipid solubility, or electronic properties, e.g. $pI_{50} = f(\text{physicochemical parameters})$

QSAR has advantages of speed and simplicity, and it can, in some cases, account for some transport and metabolic processes which occur once the compound is administered. Hence, the method is often applicable to the analysis of *in vivo* data. However classical QSAR has limitations in that it cannot handle stereoisomers, cannot correlate compounds where the base structure varies widely, and cannot implicitly handle nonlinear dependencies and interaction terms between the parameters, and QSAR analyses can be difficult to interpret in terms of mechanism at the molecular level. New QSAR methods have been developed recently which overcome some of these shortcomings. This paper discusses several of these novel molecular representations: the use of neural networks in SAR, the application of these to bioactive compound design, and the simulation of combinatorial discovery.

NEW MOLECULAR REPRESENTATIONS

Many types of molecular representation have been proposed, from Hansch parameters, (ref 1, MR, log P etc.) for chemical graph-based methods.[3−5] Recently several new representations have been devised: atomistic counts;[6] molecular eigenvalues;[7] E-state fields;[8] topological autocorrelation vectors;[9] various molecular fragment-based hash codes;[10,11] and molecular holograms.[12] These reprsentations may have advantages in speed of computation, in more accurately representing molecular properties most relevant to receptor activity, or in being more generally applicable to diverse chemical classes acting at a common receptor than the traditional representations.

Atomistic Representations. In this deceptively simple approach,[6] molecules are represented simply by counting the numbers of atoms of specific elemental type, with specific numbers of connections (a measure of the hybridization). For instance a carbon atom with four connections is denoted C4, those with three connections C3, etc., and the numbers of atoms of each type can be totaled. Figure 1 illustrates an encoding example. Although simple this representation is adequate to encode not only physicochemical parameters, such as lipophilicity and molar refractivity, but also biological activity (DHFR inhibition[6]). The fact that steric and lipophilic factors are often important in drug receptor interactions provides a partial explanation as to how such a simple representation may work.

Molecular Eigenvalues. A previous version of this eigenvalue index[l3] has been developed further by Pearlman to become the BCUT (Burden, CAS, University of Texas) index.[l4] In the present context the eigenvalue indices can

† Monash University. E-mail: Frank.R.Burden@sci.monash. edu.au.
‡ CSIRO Division of Molecular Science. E-mail: Dave.Winkler@molsci.csiro.au.

10.1021/ci980070d CCC: $18.00
Published on Web 10/13/1998

Car	C4	C3	C2	Nar	N3	N2	N1	O2	O1	S	P
6	6	4	0	0	1	1	0	0	1	0	0

Cl	F	Br	I	7-rings	6-rings	5-rings	4-rings	3-rings
1	0	0	0	1	2	0	0	0

Figure 1. Example representation of tetrazepam.

be thought of as quantifying the most electronegative and electropositive atoms in the moleculas. This comes about because the diagonal elements of the modified adjacency matrix have been ascribed atom specific values, while the off-diagonal elements have values proportional to the bond orders. The diagonalization process in effect ascribes the bond electrons back to the atoms (the trace of the matrix remaining invariant). The indices used are the 10 eigenvalues having the largest absolute magnitude and thereby represent the atoms of greatest charge.

Functional Group Representations. The other simple representation comes out of the work done by Andrews and his co-workers[15] on deducing the relative contributions of functional groups in molecules to the binding to receptors. There is some overlap with the atomistic representation of Burden, but the focus of the work was the decomposition of molecules into functional groups capable of interacting with receptors. As these functional groups were derived from 200 drug molecules of diverse structure binding to a diverse range of receptors, they are likely to be capable of general application. The functional group representations found to be statistically significant in Andrews' analyses were CO_2^-, PO_4^{2-}, N^+, N, OH, C═O, O/S ethers, halogens, Csp^3, Csp^2 and an entropic term related to the number of freely rotatable bonds in the molecule. As in Burden's atomistic approach, the molecule is simply broken down into its constituent numbers of functional groups, which are then used as a representation.

Molecular Multipole Moments. Both of the above representations are simple to implement for very large numbers of compounds with diverse structures. However, a recent paper by Platt and Silverman[16] introduced a third general representation which is intuitively appealing. They generated the zero-, first-, and second-order molecular multipole moments with respect to atomic mass and atomic charge. A previous paper[17] provided a method of defining a rotationally invariant frame of reference for the second-order mass moment (the principal moments of inertia with respect to the centre of mass) and second-order charge moment (the electric quadrupole moment with respect to the "center of dipole"). They utilized this representation to carry out 3D QSAR (the CoMMA method). This method has the advatage of accounting for more 3D properties of molecules, such as isomerism, conformation, etc. albeit at greater computational expense. We are currently working on generating analogous "lipophilic" molecular multipole moments (hydropoles), by utilizing the hydrophobic atom constant approach of Abraham and Leo[18] (see Table 1). Such hydropole moments may find application to a wide range of QSAR problems, by analogy with the application of the HINT! lipophilic fields in CoMFA.[19]

Molecular Hologram Generation. A very recent development is the molecular hologram, which is derived from a common strategy to increase the efficiency of database searching by translation of chemical structure representations into binary bit strings, known as fingerprints. Several approaches to fingerprinting have been implemented within commercial software.[20]

Table 1. Physical Interpretation of First Three Molecular Multipole Moments

order	mass	charge	lipophilicity
0	$\sum m_i$	$\sum q_i$	$\sum f_i$
1	0 (com)	μ	lipophilic dipole moment
2	moments of inertia	electrostatic quadrupole moments	lipophilic quadrupole moments

Fingerprints are generated using a combination of the keyed and hashed fingerprinting approaches. The hashed fragments encode all unique linear, branched, and cyclic fragments, including overlapping fragments. Typically fragments of size 4−7 atoms, ignoring hydrogen atoms, are generated and hashed into bins of the fingerprint. Each corresponding fragment is then mapped to a pseudorandom integer in the range $0-2^{31}$ using the CRC (cyclic redundancy check) algorithm. The integer generated by the CRC algorithm is unique and reproducible for each and every unique molecule. The hashing then occurs by folding the pseudorandom integer for a particular molecule into the bin range defined. Since the length of the bit string is considerably smaller than the integer to which the molecule is mapped, it is possible for different molecules to hash to the same bin (a "collision"). Unique fragments are always hashed into the same bin. Hologram lengths are usually prime numbers so as to minimize collisions. Rather than using a binary bit string containing either 0 or 1 in each bin, a molecular hologram retains a count of the number of times each bin is set.

The PLS technique is then used to generate a statistical model that relates the descriptor variables (occupancy numbers of the bins in the hologram) to an observable property, for example the biological activity expressed as $-\log IC_{50}$. The predictive power of the model is determined by using statistical cross-validation using a number of cross-validation groups. For the final model, the QSAR analysis is redone with the number of components set to the optimal number of components identified through the cross-validated analyses and the number of cross-validation groups set to zero. The selection of the hologram length leading to the "best" HQSAR is based on that PLS analysis that gives either the highest cross-validated r^2 or the lowest standard error associated with the cross-validation analysis. The results printed after cross-validation include the optimal number of components, which corresponds to either the maximum value of r^2 or the lowest standard error depending on the chosen criterion.

Molecular holograms, eigenvalue descriptors, molecular multipole moments, chemical graph theory, and several other developments have significantly improved the mathematical description of molecules for use in SAR studies and rational design.

Application: Holographic QSAR of Benzodiazepines. Benzodiazepines have been used therapeutically as anxiolytics, as tranquilizers, and as anticonvulsants in epilepsy. They act via the benzodiazepine site (BzR) on the γ-aminobutyric acid receptor ($GABA_A$) family and have been

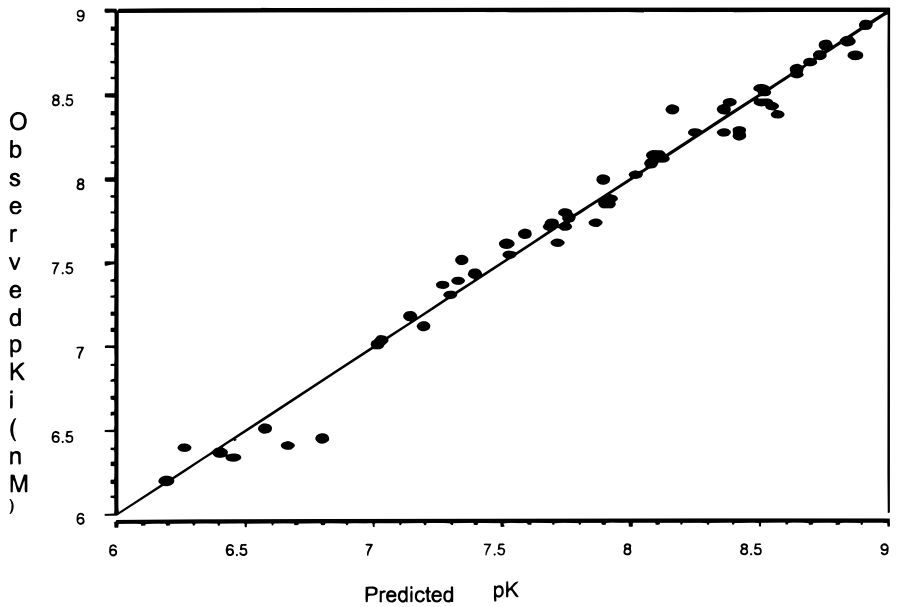
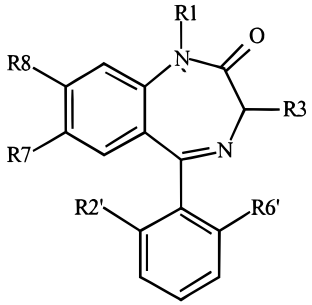


Figure 2. Observed vs predicted pI_{50} values for the 3−10 fragment size range and final (non-cross-validated) PLS model.

Scheme 1. General Benzodiazepine Structure



subject to extensive research, with over 20 QSAR studies having been carried out.[21−23] Many types of compounds have been shown to bind at the BzR e.g. benzodiazepines, arylpyrazolo-quinolines, β-carbolines, imidazo-pyridazines, and cyclopyrrolones. Structure−activity relationships in Bz receptor ligands have been reviewed recently.[24 25] We have used molecular holograms to generate QSAR models for a set of benzodiazepines active at the $GABA_A$ receptor.[12] We present a summary of the important finding on the application of molecular holograms to benzodiazepine SAR. A more detailed description is published elsewhere.[12]

The data set used for studying the application of molecular holographic representations was a set of 57 1,4-benzodiazepin-2-ones (Scheme 1) and used the HQSAR method to generate molecular representations and derive SAR models.[26] This data set has been analyzed by several groups[10,22,27,28] using a number of molecular representations and SAR mapping techniques and will allow the efficacy of the molecular hologram representation to be compared with others.

The molecular hologram representation generated by the HQSAR package (Tripos Associates) was used for this work. Figure 2 shows the excellent SAR model obtained using the largest hologram fragment size. We found the quality of the model increases as the size of fragments, and fragment complexity, increases (see Table 2). Maddalena and Johnston[27] also studied this data set and found the quality of their SAR model was dependent on the parameter set (molecular representation) used but not the architecture of the neural nets used. After pruning the input structural parameters to the optimum 10, and using a two-layer net, the cross-validated r^2 was 0.803 and the training r^2 was 0.880, with a standard error of 0.254. Changing the number of processing elements in the hidden layer made little difference to the final QSAR model. This compares with a cross-validated r^2 of 0.784 and a training $r^2 = 0.982$ with standard error = 0.119 for the best holographic QSAR model.

Table 2. Dependence of QSAR Model on Molecular Hologram Size

fragment length	hologram length	q2	std error	no. of components
1	53	0.381	0.589	3
1−3	83	0.592	0.479	3
4−7	199	0.701	0.460	14
3−10	59	0.784	0.411	18

So and Karplus[28] also analyzed the same benzodiazepine data using a genetic neural net. This approach uses a genetic algorithm to preselect the descriptors and a neural net to map structure to activity. They achieved better QSAR models using this approach. They raised the question as to whether the selection of descriptors used by Maddalena and Johnston were optimal. So and Karplus' best model, C10-3, had a training r^2 of 0.941 and a cross-validated r^2 of 0.882.

Greco and co-workers analyzed this data set using classical QSAR, with descriptors derived from molecular orbital calculations, and molecular field analysis (CoMFA).[22] With MO descriptors their models had low cross-validated r^2 values (0.12−0.46) unless indicator variables were used. The models derived from CoMFA analyses using electrostatic alone, or steric and electrostatic fields also gave relatively low cross-validated statistics. When they used the $^1\chi_7$ descriptor to account for lipophilicity of the substituent at position 7, the models improved. The best model had a cross-validated r^2 value of 0.76 with six latent variables and a standard error of 0.41.

In a recent study[10] we used atomistic and functional group descriptors to represent structures, with PLS and several

Table 3. Comparison between Benzodiazepine QSAR Models

	training		cross-validated	
study	r^2	std error	r^2	std error
Maddalena & Johnston 1995 NN	0.880	0.254	0.803	0.321
So & Karplus 1996 NN	0.941	not reported	0.882	not reported
Winkler et al. 1998a NN	not reported	not reported	0.824	0.315
Winkler et al. 1998a PLS[a]	not reported	not reported	0.331	0.608
Winkler et al. 1998b (size 1[b]) PLS	0.543	0.506	0.381	0.589
Winkler et al. 1998b (size 1−3) PLS	0.715	0.399	0.592	0.479
Winkler et al. 1998b (size 4−7) PLS	0.965	0.156	0.701	0.460
Winkler et al. 1998b (size 3−10)	0.982	0.119	0.784	0.411

[a] PLS study with five latent variables and 54 compounds. [b] Size refers to hologram fragment size range.

neural net architectures to develop the SAR models. Only three layer neural nets were used and, unlike Maddalena and Johnson, we found considerable variation in the quality of the SAR and the architecture of the neural net. Our optimum model, a positionally-independent atomistic representation, gave a cross-validated r^2 of 0.824 and a standard error of 0.315. We also analyzed the data using PLS and obtained a lower cross-validated r^2 of 0.331 with standard error of 0.608. This is similar to the statistics from the molecular fragment size 1 holographic QSAR study, which had a cross-validated r^2 of 0.381 and a standard error of 0.589. This suggests the unit length molecular hologram and the atomistic representation we used are encoding similar information.

Inspection of Table 3 shows that the QSAR models from neural net studies give much better cross-validated statistics (r^2 values of 0.8−0.9) than do PLS studies (0.3−0.7). This suggests that linear regression methods are not capable of accounting for interactions between parameters and nonlinearity in the structure−activity response surface. The lower performance of linear techniques such as least squares regression compared with a nonlinear technique such as an ANN suggests that significant nonlinearity exists in the Bz data set. These nonlinearities may relate to higher order representations implicit in the data. Use of molecular holograms with ANNs is likely to yield even more accurate models based on the experience with other studies using ANN with traditional substituent constants as molecular representations (Table 3). While the representation used is simple, it is still an intrinsically more "information rich" representation than other indices, such as that of Randić and Trinajstić,[5] where all molecular properties are compressed into a single value. Molecular hologram representations encode implicit information relevant to receptor−ligand interactions. It appears that the complex inter-relationships between elements of simple molecular fragment representations may be capable of encoding subtle molecular structural properties. This concept has been discussed recently.[8,11]

NEW SAR MAPPING METHODS: NEURAL NETWORKS

The relationship between the molecular structure and the biological activity is modeled using techniques such as Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares (PLS),[29] and artificial neural networks (ANN).[30−33] As we foreshadowed above, in many instances ANNs have proven to be better than MLR, PCR, or PLS.

Artificial neural networks are computer-based mathematical models developed to have analogous functions to idealized simple biological nervous systems. They consist of layers of processing elements (neurodes), which are considered to be analogous to the nerve cells (neurons), and these are interconnected to form a network which is in essence a parallel computer[30] even though they are most likely to be run on nonparallel computers such as personal computers or workstations.

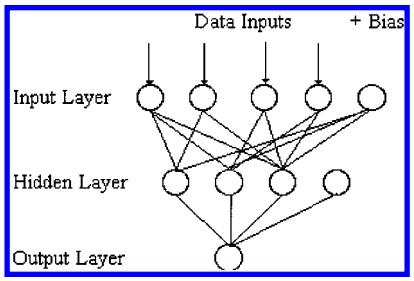


Figure 3. Typical back propagation neural net architecture.

There are two main classes of ANN, supervised and unsupervised, and several types within each class. Supervised ANNs are trained to build internal algorithms relating patterns of inputs to outputs. After learning the relationship between the inputs and outputs they are able to classify patterns and make decisions or predictions based upon new patterns of inputs. The backpropagation neural net will be discussed here, as it is the type most widely applied to QSAR problems (Figure 3).[32]

The nonlinear output from each neuron is considered to be another important feature of ANNs. If many neurons each with nonlinear inputs and outputs are interconnected, then the network essentially becomes a universal mathematical function approximator.[31−33] Consequently, ANNs are useful in predicting results from complex data interactions involving multiple variables where the inter-relationships among them are poorly understood or fuzzy.

Model Validation: Bayesian Regularization. We have recently investigated the use of Bayesian regularization in artificial neural nets.[12] Using Bayesian regularization[34] removes the need to supply a validation set since it minimizes a linear combination of squared errors and weights. It also modifies the linear combination so that at the end of training the resulting network has good generalization qualities. It has also been suggested that there is no need for a testing set since the application of the Bayesian statistics provides a network that has maximum generalization. Our study used a network architecture with three hidden nodes which proved to be more than sufficient in all cases with the Bayesian regularization method estimating the number of effective parameters. The concerns about overfitting and overtraining

are also removed by this method so that the production of a definitive and reproducible model is attained. There remains a minor problem of instability which is caused by the provision of randomized weights at the start of training. Some such randomized sets can cause the training to produce near zero final weights; this can be overcome by training a number of nets with different starting weights and selecting the best standard error of prediction. It has been found the those networks that converge to finite weights always produce near identical answers.

Application: Simulation of Combinatorial Discovery. The development of combinatorial chemistry, and the resultant large increases in the numbers of chemical entities screened for drug activity, has resulted in a paradigm shift in the way new drug leads are discovered. It is now possible to generate millions of chemical analogues in a relatively short time with greatly reduced effort. Rapid screening techniques have necessarily emerged to keep pace with the generation of new combinatorial libraries. It is now routine to carry out 40−50 000 screening events per week with a small number of staff.[35] Recombinant technologies have contributed to this scaling up of screening throughput by allowing ready access to quantities of pure receptor or protein gene products.

As powerful as these new methods of combinatorial and mass screening are, they are still only capable of accessing a very small region of the "universe" of possible chemistries. Simple consideration of the numbers of possible branched chain isomers of alkanes shows that there are over 4 billion C_{30} isomers alone. Clearly more complex compounds with heteroatoms, rings, and unsaturation are capable of being assembled in an almost infinite variety of ways. Estimates of the "universe" of chemical compounds that it is possible to synthesize by combinatorial methods range from 10^{60}−10^{400}, numbers so vast that only a minute fraction could conceivably be generated and tested by combinatorial methods. This recognition is driving the quest for methods of simulation of combinatorial synthesis and high throughput screening "*in silico*". Methods to allow exploration of much larger region of combinatorial space would be of considerable interest in allowing a focusing of the combinatorial chemistry effort into chemical species with inherent novelty and receptor efficacy.

Recently, Ho and Marshall[36] described a technique for generating very large databases, representing a "virtual" combinatorial library, using a procedure they called DB-Maker. This involved permutation of SMILES strings coupled to 3D structure building methods such as Concord,[37] to generate many chemically feasible 3D structures. Tripos Associates have used an alternative approach which exploits simple chemistries and commercially-available building blocks to generate a 3D database. This ChemSpace database contains approximately 1 *trillion* chemical structures for use in similarity and pharmacophore searches, approximately 50 000 times more than all the compounds in CAS.

We have utilized the concept of a QSAR model as a "virtual receptor" to allow rapid screening of these "virtual combinatorial libraries". Previously we, like others, have investigated the use of neural networks and novel molecular representations in QSAR studies. We are working toward the development of computationally cheap, simple molecular representations for use in these studies involving large data sets which, coupled with developments in neural networks, will facilitate the generation of receptor surrogates with useful properties. Our implementation of virtual receptors involves a trained neural network and simple molecular representations which would be capable of rapidly evaluating large numbers of compounds for possible activity against the receptor type. Such virtual receptors would be useful screening paradigms for finding leads in large virtual databases. We have investigated this possibility using ANNs and found the approach feasible. Related work by Maddalena and Johnston[27] has given support to these concepts in showing that ANNs could provide a pharmacophore model with receptor-like properties. More recently a number of different approaches to the library design and virtual screening have been reported (e.g. refs 38−43). Tropsha and his group have developed a method (Focus-2D) for searching virtual libraries for structures similar to biologically active compounds using simulated annealing and topological descriptors.[38] Shi et al. have used genetic function approximations to carry out QSAR studies in the NCI database which describe antitumour activity patterns.[39] Screening of virtual libraries using 3D steric and electronic grids has been reported by Lui et al.[40] Horvath[41] has automated the conformational analysis and active site docking of a 2500 library of potential trypanothione reductase inhibitors. Vedani, Dobler, and Zbinden[42] have developed a quasi atomistic receptor model for use in screening of libraries which defines a pseudoreceptor surface with properties which adapt to the requirements of the traing set. Polanski has used a self-organizing map to derive a receptor-like neural network which could be used to screen virtual libraries. A flexible pharmacophore model of another receptor type has recently been described using a genetic algorithm approach.[44]

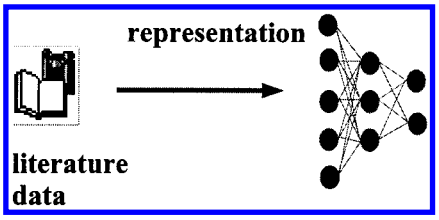


Figure 4. Virtual receptor modeling using neural networks.

The problem of defining a virtual receptor, once a suitable set of measurements of biological activity of compounds at the receptor is compiled, involves generation of a suitable molecular representation for the compounds whose activity have been determined and mapping of molecular representation to biological activity (Figure 4).

Benzodiazepine Virtual Receptor. We applied some of the new fragment-based molecular representations and neural networks to the concept of virtual receptors. We used two simple representations: the atomistic representation,[6] which superficially appears to disregard much information such as topology and stereochemistry; and the functional group-based representation.[15]

A data set was compiled from the literature.[28,45−53] It consisted of 321 compounds of diverse structure: benzodiazepines, arylpyrazolo-quinolines, β-carbolines, imidazopyridazines, and cyclopyrrolones. These were broken up into two sets: 21 compounds would serve as the test set, whilst the other 300 compounds would form the basis of training and validation sets. Training sets consisted of 270 com-

Table 4. Comparison of Neural Network and MLR

method	RMSE (Val)	RMSE (test)
MLR	0.158 (0.852)[a]	0.150 (0.809)
ANN 21:6:3:2:1	0.137 (0.739)	0.147 (0.793)
ANN 21:8:5:3:1	0.101 (0.545)	0.119 (0.642)

[a] Unscaled data.

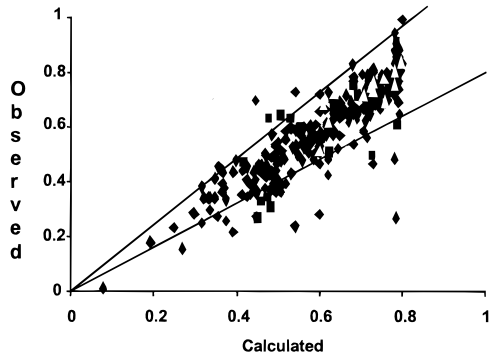


Figure 5. Output from a 21:8:5:3:1 network.

pounds; validation sets consisted of 30 compounds. Thus, cross-validation involved the generation of 10 training and validation set pairs. The neural network produced in each case was tested using the test set. It is worth briefly explaining the difference between a validation and test set. The validation set is used to determine the set of weights which give the best predictions. The test set was then tested on those weights, to provide further verification of the neural networks predictive abilities.

The representation used was based on the "atomistic" approach described previously. However, input parameters relating to the number and type of rings were added, thus affording the neural network some insight into the molecule's topology. Twenty-one input variables were used to represent each molecule: C(aromatic), C4, C3, C2, N(aromatic), N3, N2, N1, O2, O1, S, P, Cl, F, Br, I, seven-membered rings, six-membered rings, five-membered rings, four-membered rings, and three-membered rings. Efficacy at the M_1 receptor was reported as log $1/IC_{50}$ (pI_{50}) for displacement of tritiated diazepam.

The ANN's used were three or four layer fully connected, feed forward networks which were trained by the use of back propagation. The order of the compounds was randomized, to ensure that when the data set was split into validation and training sets, the validation set would be truly representative of the training set. The data set was then split into a series of training and validation sets, for use in cross-validation. Care was taken to keep the network sufficiently small, in terms of the number of weights to be computed, so that overtraining was unlikely to occur.

The standard error of predictions (SEPs) using the various representations are shown in Table 4. Numerous ANN architectures were tested; the network with the lowest cross-validated SEP was deemed to be the optimal architecture. As the table shows SEPs obtained using ANNs are generally significantly lower than those obtained using the linear techniques. Rather surprisingly, the model does not suffer when positional information is removed from the representation (i.e. the position of substitution is ignored). Indeed, the best model using the atomistic approach was positionally-independent. The model obtained using the atomistic representation provides an SEP comparable to the model using the functional group representation.

The five-layer 21:8:5:3:1 network proved to be most successful at modeling the data. It had the lowest RMSE for the whole data set and a correlation coefficient of $R = 0.794$. It also had the lowest predictive RMSE, indicating that its ability to model the activity of compounds in both the validation and test sets was superior to any other architecture. In general, the results were very good with all architectures; the RMSEs were lower than one pI_{50} unit, with the 21:8:5:3:1 network having a RMSE of 0.758. A sample output from a 21:8:5:3:1 neural network is shown in Figure 5. An indication of how successful the neural network has been in modeling the data set comes in comparing the above results with those obtained using MLR. The results are summarized in Table 4.

The resulting virtual receptor was used to screen a virtual combinatorial library simulated by 40 000 compounds from the Maybridge chemical database. Several compounds were predicted to have high affinity for the benzodiazepine receptor, and their activity is currently under investigation.

REFERENCES AND NOTES

(1) Rouvray, D. H. *New Scientist* **1993a**, May 29, 35−38.
(2) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**. *86*, 1616.
(3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis;* J. Wiley and Sons: New York, 1986.
(4) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(5) Randić, M.; Trinajstić, N. *J. Mol. Struct.* **1993**, *300*, 551−571.
(6) Burden, F. R. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7−11.
(7) Burden, F. R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309−314.
(8) Kier, L. B.; Hall, L. H. In *Reviews in Computational Chemistry;* Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 2, p 374.
(9) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Chem. Inf: Comput. Sci.* **1996**, *36*, 1205−1213.
(10) Winkler, D. A.; Burden, F. R.; Watkins, A. *Quant. Struct.-Activ. Relat.* **1998**, *17*, 14−19.
(11) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(12) Winkler, D. A.; Burden, F. R. *Quant. Struct.-Activ. Relat.* **1998**, *17*, 224−231.
(13) Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−27.
(14) Pearlman, R. S.; Stewart, E. L.; Smith, K. M.; Balducci, R. Novel Software Tools for Combinatorial Chemistry and Chemical Diversity. Paper given at the 1997. Charleston Conference *Advancing New Lead Discovery*, Isle of Palms, SC (March 1997).
(15) Andrews, P. R.; Craik, D. J.; Martin, J. L. *J. Med. Chem.* **1984**, *27*, 1648−57.
(16) Platt, D. E.; Silverman, B. D. *J. Comput. Chem.* **1996**, *17*, 358−66.
(17) Silverman, B. D; Platt, D. R. *J. Med. Chem.* **1996**, *39*, 2129−40.
(18) Abraham, D. J.; Leo, A. J. *Proteins: Struct. Funct. Genetics* **1987**, *2*, 130−152.
(19) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−552.
(20) Tripos Associates, 1699 South Hanley Road, Suite 303, St. Louis, MO 63144.HQSAR Software v 1.0. Tripos Associates: (http://www.tripos.com/products/hqsar.html).
(21) Blair, T.; Webb, G. A. *J. Med. Chem.* **1977**, *20*, 1206−10.
(22) Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. *Quant. Struct.-Act. Relat.* **1992**, *11*, 461−77.
(23) Gupta, S. P.; Paleti, A. *Quant. Struct.-Act. Relat.* **1996**, *15*, 12−16.
(24) Gupta, S. P. *Chem. Rev.* **1989**, *89*, 1765−1800.
(25) Villar, H. O.; Davies, M. F.; Loew, G. H.; Maquire, P. A. *Life Sci.* **1991**, *48*, 593−602.
(26) Haefely, W.; Kyburz, E.; Gerecke, M.; Mšhler, H. *Adv. Drug Res.* **1985**, *14*, 165−322.
(27) Maddalena, D.; Johnston, G. A. R. *J. Med. Chem.* **1995**, *38*, 2824−2836.
(28) So, S.-S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 5246−5256.
(29) Brereton, R. G. *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems;* Ellis Horwood: New York, 1990.

(30) Rumelhart, D. E.; McClelland, J. L. *Parallel distributed processing: Explorations in the microstructure of cognition;* MIT Press: Cambridge, 1986; Vols. I and II.
(31) Salt, D. W.; Yildiz, N.; Livingstone, D. J.; Tinsley, C. *Pestic. Sci.* **1992**, *36*, 161−170.
(32) Hornik, K.; Stinchcombe, M.; White, H. *Neural Networks* **1988**, *2* 359−366.
(33) Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1229−1231.
(34) MacKay, D. J. C. A Practical Bayesian Framework for Backprop Networks. *Neural Computation* **1992**, *4*, 415−447.
(35) Rouvray, D. H. *Chem. Brit.* **1993b**, June, 495−498.
(36) Ho, C. M. W.; Marshall, G. R. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 65−86.
(37) Rusinko, A.; Sheridan, R. P.; Nilakanta, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1989**, *29* 251−255.
(38) Zheng, W.; Cho, S. J.; Tropsha, A. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 251−58.
(39) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 189−99.
(40) Lui, D.; Jiang, H.; Chen, K.; Ji, R. *J. Chem. Inf. Comput. Sci* **1998**, *38*, 233−42.
(41) Horvath, D. *J. Med. Chem.* **1997**, *40*, 2412−23.
(42) Vedani, A.; Dobler, M.; Zbinden, P. *J. Am. Chem. Soc.* **1998**, *120*, 4471−77.
(43) Polanski, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 553−61.
(44) Walters, D. E.; Hinds, R. M. *J. Med. Chem.* **1994**, *37*, 2527−36.
(45) Zhang, W.; Koehler, K. F.; Harris, B.; Skolnick, P.; Cook, J. M. *J. Med. Chem.* **1994**, *37*, 745−757.
(46) Harrison, P. W. *Eur. J. Med. Chem.* **1996**, *31*, 651−662.
(47) Davies, L. P.; Barlin, G. B.; Ireland, S. J.; Ngu, M. M. L. *Biochem. Pharmacol.* **1992**, *44*, 1555−1561.
(48) Barlin, G. B.; Davies, L. P.; Davis, R. A.; Harrison, P. W. *Aust. J. Chem.* **1994**, *47*, 2001−2012.
(49) Fryer, R. I.; Zhang, P.; Rios, R.; Gu, Z.-Q.; Basile, A. S.; Skolnick, P. *J. Med. Chem.* **1993**, *36*, 1669−1673.
(50) Wang, C.-G.; Langer, T.; Kamath, P. G.; Gu, Z.-Q.; Skolnick, P.; Fryer, R. I. *J. Med. Chem.* **1995**, *38*, 950−957.
(51) Hollinshead, S. P.; Trudell, M. L.; Skolnick, P.; Cook, J. M. *J. Med. Chem.* **1990**, *33*, 1062−1069.
(52) Allen, M. S.; Hagen, T. J.; Trudell, M. L.; Codding, P. W.; Skolnick, P.; Cook, J. M. *J. Med. Chem.* **1988**, *31*, 1854−1861.
(53) Yokoyama, N.; Ritter, B.; Neubert, A. D. *J. Med. Chem.* **1982**, *25*, 337−339.

CI980070D