The Applications of Artificial Neural Networks in the Identification of Quantitative Structure-Activity Relationships for Chemotherapeutic Drug Carcinogenicity

Alexander C. Priest, Alexander J. Williamson, and Hugh M. Cartwright

Department of Chemistry, University of Oxford
Physical & Theoretical Chemistry Lab, South Parks Road, Oxford, UK, OX1 3QZ
alexander.priest@chem.ox.ac.uk

Abstract. We investigate which of two Artificial Intelligence techniques is superior at making predictions about complex carcinogen systems. Artificial Neural Networks are shown to provide good predictions of carcinogen toxicology bands for drugs which are themselves used to treat cancerous cells, by using a novel system of molecular descriptors derived from the molecules' mass spectrometry intensities, reduced in dimensionality by Principal Component Analysis, to form a series of orthogonal descriptors which retain 95% of the variance of the original data.

The creation of molecular descriptors from PCA-resolved mass spectrometry data is shown to be superior to the use of Self-Organising Maps, the selection of a series of modal fragments, or the use of every peak (within the confines of the precepts of Artificial Intelligence). A new system of backpropagation which increases network efficacy in this case is also proposed.

Keywords: artificial neural network, ANN, carcinogenicity, QSAR.

1 Introduction

Computer-based drug design is an important area of pharmaceutical chemistry [1,2]. Quantitative Structure-Activity Relationships (QSARs), determined computationally from experimental observations, are widely used to suggest candidate drugs for early screening, reducing both the time and money spent on synthesis and *in vivo* testing [3]. Computational approaches to the design of new drugs have become progressively more significant and sophisticated as computer power has increased, and all major pharmaceutical companies now make extensive use of these methods.

Many properties of a chemical contribute to the decision as to whether it is viable as a drug. Apart from the obvious question of efficacy, the harm that a drug itself presents must be sufficiently low that it not outweigh the drug's therapeutic value. The carcinogenicity and teratogenicity of a drug is an important area in this regard, but modelling these kinds of complex and dynamic systems presents a significant barrier to the advancement of this technology. Chicu, at al. [4] have used statistical methods to predict chemicals' toxicity in crustacea and fish, while Williams [5]

P.R. Cohen, N.M. Adams, and M.R. Berthold (Eds.): IDA 2010, LNCS 6065, pp. 137–146, 2010.

predicted the carcinogenic effects of ATPase inhibitors using chemical structure. Conolly et al. [6] modelled respiratory tract carcinogens in a similar way.

Artificial Intelligence techniques have, with great success, provided insight into complex data in other spheres [7,8,9,10,11,12,13,14] suggesting that techniques such as Artificial Neural Networks (ANNs) will provide an accurate link between molecular descriptors and physical data.

However, numerous potential architectures, parameter settings and dimensionality-reduction techniques exist in Artificial Neural Networking and optimising these is an essential step in the use of the method. Furthermore, it is necessary to represent physical phenomena in a theoretical frame and, consequently, discerning a representative series of molecular descriptors is imperative. Matter [15] and Hopfinger et al. [16] proposed that this could be achieved by creating a series of molecular descriptors to represent an energy-minimised three-dimensional drug, and Tanabe et al. [17] subsequently predicted carcinogenicities using three-dimensional geometrical molecular descriptors as a basis for supervised learning. We describe here an extension of their work using mass spectral data, which is surprisingly effective considering its relative simplicity.

2 Molecular Descriptors and Data

A molecular descriptor specifies some feature of a molecule, such as its dipole moment, size, or electron distribution, in an unambiguous way; the complete description of a molecule thus requires many descriptors. Since we can anticipate that the effectiveness of drugs must be related in some way to properties at the molecular level, it is evident that connections exist between a suitably-chosen set of molecular descriptors and the activity of a drug.

Although the most widely used molecular descriptors are a direct expression of properties of the molecule itself, there are many examples where some derivable property of the molecule, such as its infrared spectrum, have been used as descriptors: this is a standard approach in Computational QSAR analysis. We have therefore investigated whether the mass spectra of candidate drugs might constitute suitable proxy molecular descriptors. As mass spectra contain substantial amounts of information about a sample, it is reasonable to suppose that a single mass spectrum might be able to replace a substantial number of more conventional descriptors. Mass Spectrometry data were obtained from the National Institute of Standards and Technology. The data series included a CAS Number, and a list of peaks and their intensities.

The wealth of data in a mass spectrum presents, however, a problem as well as an advantage. When seeking correlations between an experimental spectrum and a drug's properties, unsupportable and unjustified interaction models might emerge through overfitting, so our initial step is to reduce the dimensionality of the data. All mass spectra were treated with Principal Component Analysis, creating a series of quasi-molecular descriptors which contained 95% of the original data variance (Table 1). The number of quasi-molecular descriptors was reduced from several thousand to 146. These new descriptors were preliminarily compared with modal (intensities of popular fragments) and banded (similarly-sized fragments banded together) peaks. These alternative systems were good comparators as they represent other valid

Table 1. 146 Principal Components representing 95% of the variance of the original data

Transformed Data	*Cumulative Responsibility*
PC 1	0.0709
PC 2	0.1303
PC 3	0.1738
PC 4	0.2076
PC 5	0.2355
PC 6	0.2596
PC 7	0.2823
…	…
PC 140	0.9460
PC 141	0.9468
PC 142	0.9476
PC 143	0.9483
PC 144	0.9491
PC 145	0.9498
PC 146	0.9504

methods of reducing the number of inputs, which is necessary in Artificial Neural Networking to generate realistic, short processing runs. However, these alternative molecular descriptors generated predictions no better than those generated through randomisation (ie if bands had been selected at random). See Section 5.

2.1 Toxicity Data

Training data (oral, rat, TD_{50}) were sourced from the National Toxicology Program (NTP) [18] and The Carcinogenic Potency Project, Berkeley University [19], and carefully concatenated and their units reconciled. These data were logarithmically banded into nine distinct groups (1-9; 1=most carcinogenic, 9=least carcinogenic), and tabulated along with the CAS Number of the chemical which they represented.

3 Self-Organising Maps

The PCA-resolved mass spectra molecular descriptors were used as inputs for 12×12 hexagonal self-organising maps (SOMs) with various interaction settings. No simple relationship between a 2-dimensionally-mapped series of molecular descriptors and the carcinogenicity of a molecule was found (see section 5). Following a series of runs, using a wide range of parameters and architectures this method of dimensionality reduction was discounted. Computational physical and biological scientists should be wary of the use of generic techniques in data mining, except on a case-by-case basis with proper double cross-validation testing. It is the nature of complex systems, such as the activity of carcinogens, that it is hard to understand in advance why one technique succeeds where another fails in discerning patterns and making predictions.

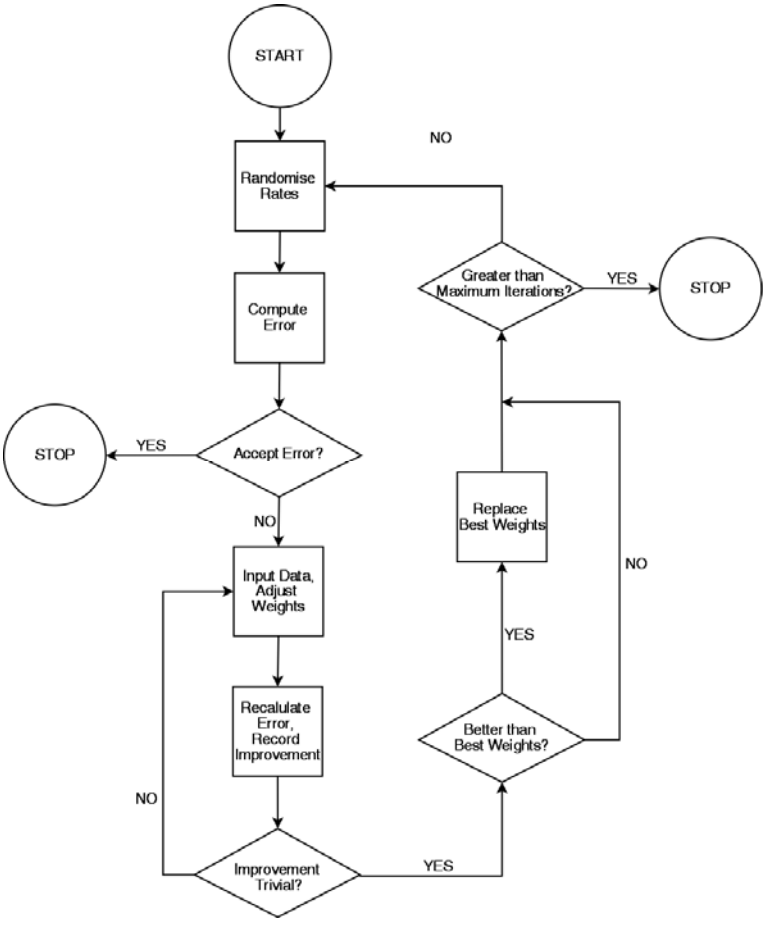

Fig. 1. The random-start Self-Organising Map Decision Tree

4 Artificial Neural Networks

Standard feed-forward Artificial Neural Networks, using a sigmoidal transfer function and a novel backpropagation system (which also guarded against pareto equilibria, when a Neural Network falls into a false minimum, from which it is unable to escape), were then applied to the problem. Where the network was moving away from a real equilibrium a randomisation of the weights occurred, and processing began again. The new system of backpropagation works as follows:

Let w = the weight assigned to the connection between two nodes

O = output value of a node, following application of the transfer function

$\varepsilon = T - O$ = the overall error (where T is the target value)

Then the error of the output node,

$$\varepsilon_{kj} = \psi^{-1}(O_{kj}) \times \psi^{-1}(T_{kj}) \tag{1}$$

where $T_{output,j}$ = the ideal output, and the inverse transfer function (reverse of the TF),

$$\psi^{-1}(t) = 100 \times \log\left(\frac{1+\frac{t}{2000}}{1-\frac{t}{2000}}\right) \tag{2}$$

Therefore, the change in connection weight,

$$\Delta w_{kj} = \ell \varepsilon_{kj} \frac{O_{k\text{-}1,j} \times w_{k,j}}{\sum_i O_{k-1,i} \times w_{k,i}} \tag{3}$$

where ℓ is the learning parameter. Therefore, let the target output of the hidden layer,

$$T_{kj} = \frac{\sum_i \frac{T_{k+1,i} \times O_{kj} \times w_{k+1,j}}{\sum_h O_{kh} \times w_{k+1,h}}}{N_{k+1}} \tag{4}$$

where N_k=number of nodes in layer k. The system is then processed backwards.

The new system stems from the necessity, with a more complicated data set, to invoke a similar change in error during backpropagation, which will prevent overcorrection. This decreases the time a system takes to arrive at a minimum. This is achieved by calculating the *inverse* error gradient[1] at each node and then using this to calculate the amount by which the weights should be altered in the subsequent iteration. This allows the network to spring out of pareto equilibria when the amount by which each weight changes in each iteration is high, while not adversely affecting the final equilibrium because the change increment is small towards the end. This system was found to reduce typical processing time by 11% as fewer iterations are necessary to achieve the same minimum equilibrium[2].

---

[1] The inverse error gradient is the reciprocal of the error gradient, ie $\left.\frac{\partial \varepsilon}{\partial w_i}\right|_{w_{i \neq j}}$.

[2] That is to say that over seven program runs using different carcinogen data sets, the novel system of backpropagation arrived at the minimum error on average 11% faster than an equivalent network using standard backpropagation.

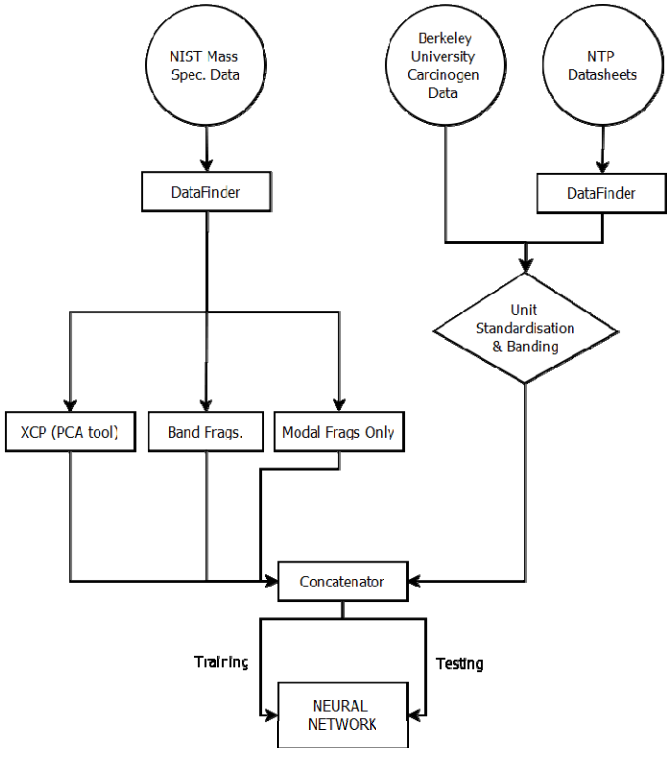

Fig. 2. A data flow diagram for the Artificial Neural Network Method

5 Results

Through the use of ANNs, we were able to accurately identify a QSAR for the prediction of carcinogenicity using PCA-resolved Mass Spectral data to describe the molecules (Fig 3). This QSAR was tested using a series of molecules not present in the training set and also by the leave-three-out cross validation method. Leave-three-out was selected as opposed to leave-one-out to ensure that the system was resolving in favour of dissimilar molecules. 174 out of 196 test molecules were banded to within the success criterion (within one band of their actual band as given by Berkeley or the NTP), representing an accuracy rate of 89%, which we consider viable. This compares to 53 out of 196 selected accurately following the use of Self-Organising Maps.

Table 2. Average % of correctly-identified carcinogens (one band margin of error)

Processing Architecture	% Accuracy (average)	Weighted-kappa Value
Banded Mass Spec + ANN	20	0.18
Modal Mass Spec + ANN	25	0.21
PCA Mass Spec + SOM	27	0.20
PCA-resolved Mass Spec +ANN	89	0.76

Systems in which the raw data consisted of modal (49/196) or banded (39/196) Mass Spectra failed to identify sufficiently accurate models.

As an additional, if somewhat qualitative, test of reliability, Table 3 shows predicted carcinogenicity band for some drugs compared with their International Agency for Research on Cancer classification. This classification ranks chemicals into 5 groups with the following definitions:

1: *carcinogenic to humans.*
2A: *probably carcinogenic to humans.*
2B: *possibly carcinogenic to humans.*
3: *not classifiable as to its carcinogenicity to humans.*
4: *probably not carcinogenic to humans.*

Table 3. Predicted carcinogenicity band (ANN) for some example drugs compared with their International Agency for Research on Cancer's Classification

Chemotherapeutic Drug	CAS No.	Predicted Carcinogen Band	IARC Group
Chlorambucil	305-03-3	2	1
Cyclophosphamide	50-18-0	2	1
Mechlorethamine	51-75-2	3	2A
Methotrexate	59-05-2	7	3
Fluorouracil	51-21-8	6	3
Uracil	66-22-8	4	2B

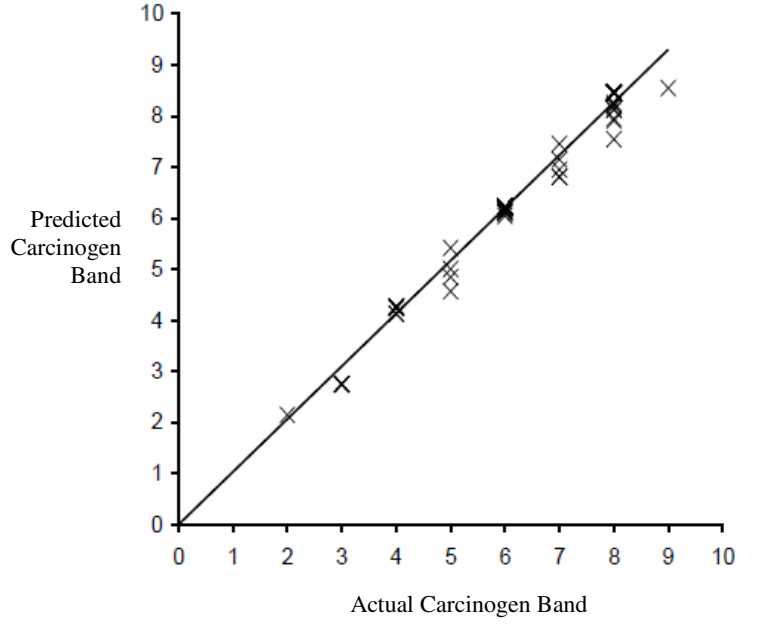

Fig. 3. Carcinogen Band prediction by Artificial Neural Network

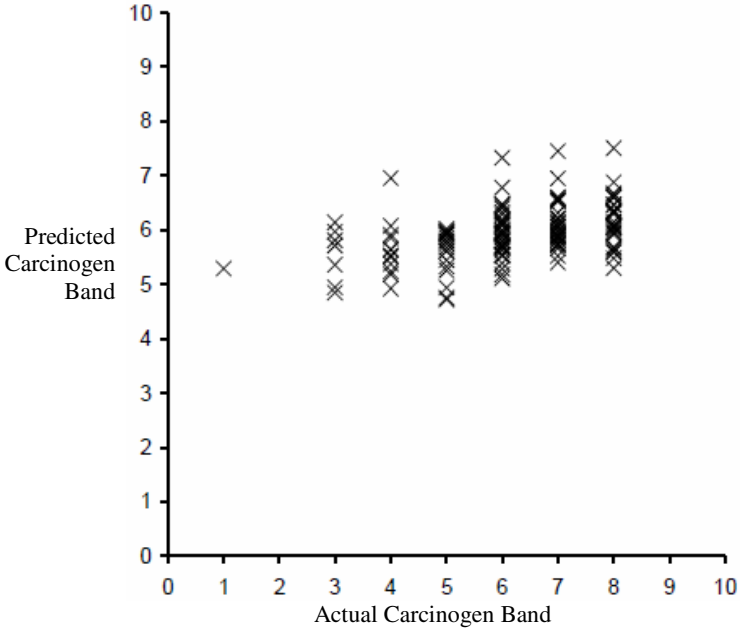

Fig. 4. Carcinogen Band prediction by Self-Organising Map

6 Conclusions

The carcinogenicity of a potential cancer drug is discernible using a series of Artificial Neural Networks and PCA-resolved Mass Spectral data. This clearly implies that mass spectral intensities may be used as reliable molecular descriptors in this, and possibly other cases, through the use of this technique. The selection of a set of the most frequent or banded fragments does not allow accurate prediction.

It has also been shown that reducing the burden of dimensionality does not necessarily compromise the reliability of the process: the selection of a smaller set of orthogonal descriptors is accurately achieved using Principal Component Analysis, but this process is not satisfactory when using Self-Organising Maps, which, with this dataset, destroy the underlying nuances of the data which act as molecular descriptors.

The novel system of backpropagation proposed is both more efficient and reliable in monitoring a series of connection weights within an Artificial Neural Network (11% faster for similar levels of accuracy), and has the potential to be applied more generally in this field.

This application of Artificial Intelligence to a complex pharmaceutical problem has wide-reaching implications for the analysis of complex biological systems by physical and life scientists, as it lays the building blocks for drug design to become even more efficient, and allows doctors prescribing nascent or experimental treatments to understand more fully the drug regime they are recommending.

Further optimisation of the network architecture and the discovery of additional data (both mass spectral and carcinogenicity) would lead to improvements in the accuracy of the network.

Various options are open as to how to proceed. An investigation into the incorporation of Support Vector Machines or Fuzzy processes into the system would be extremely interesting, as would the analysis of a Fuzzy success criterion (currently a prediction is right or wrong, but information on how wrong would be useful).

Acknowledgments

This research has been supported by St John's College, Oxford, the Department of Chemistry at Oxford University and SCAST.

References

1. Walters, W.P., Stahl, M.T., Murcko, M.A.: Virtual screening - an overview. Drug Discovery Today 3(4), 160–178 (1998)
2. Jorgensen, W.L.: The many roles of computation in drug discovery. Science 303(5665), 1813–1818 (2004)
3. Agatonovic-Kustrin, S., Beresford, R.: Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Pharm. Biomedical Analysis 22(5), 717–727 (2000)
4. Chicu, S., Hermann, K., Berking, S.: An Approach to Calculate the Toxicity of Simple Organic Molecules on the Basis of QSAR Analysis in Hydraactina echinata Hydrozoa. Cnidaria. Quant. Struct.-Act. Relat. 19 (2000)
5. Williams, W.R.: Relative molecular similarity in selected chemical carcinogens and the nucleoside triphosphate chain. Pharmacology & Toxicology 92(2), 57–63 (2003)
6. Conolly, R.B., et al.: Human respiratory tract cancer risks of inhaled formaldehyde: Dose-response predictions derived from biologically-motivated computational modeling of a combined rodent and human dataset. Toxicological Sciences 82(1), 279–296 (2004)
7. Werbos, P.J.: The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting. Wiley, Chichester (1994, a reprinting of his Harvard DPhil thesis of 1974)
8. McCulloch, W., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133 (1943)
9. De Wolf, E., Francl, L.: Neural Networks that distinguish infection periods of wheat tan spot in an outdoor environment. Phytopathology 87, 83 (1997)
10. Rayudu, R., Samarasinge, S.: A network of neural nets to model power system fault diagnosis. In: Proceedings of the Fourth International Conference on Neural Information Processing (1997)
11. Jiang, D., et al.: Progress in developing an ANN model for air pollution index forecasting. Atmospheric Environment 38, 7055 (2004)
12. Rajanayake, C., et al.: Solving the inverse problem in stochastic groundwater modelling with artificial neural networks. In: Proceedings of 1st Biennial Congress of the International Environmental Modelling Society (2002)
13. Chandraratne, M.R., et al.: Prediction of lamb tenderness using image surface texture analysis. Journal of Food Engineering (2005)
14. Limsombunchai, V., Samarasinge, S.: House price prediction: hedonic price model vs artificial neural networks. Kasetsar University Journal of Economics (2005)

15. Matter, H.: Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. Journal of Medicinal Chemistry 40(8), 1219–1229 (1997)
16. Hopfinger, A.J., et al.: Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. Journal of the American Chemical Society 119(43), 10509–10524 (1997)
17. Tanabe, K., Ohmori, N., Ono, S., Suzuki, T.: Neural network based QSARs of chemical carcinogens derived from chemical safety database CAESAR. Pharm. & Phar. 58(Suppl. 1), A32 (2006)
18. National Toxicology Program, `http://ntp-server.niehs.nih.gov/`
19. Carcinogenic Potency Project, Berkeley University, `http://potency.berkeley.edu/`