33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Predicting Odorant Chemical Class from Odorant Descriptor Values With an Assembly of Multi-Layer Perceptrons

Luqman R. Bachtiar, Charles P. Unsworth, *Member IEEE*, Richard D. Newcomb, Edmund J. Crampin

Abstract— **Chemical descriptors are a way to define information concerning the physical, chemical and biological properties of a chemical compound. Machine learning methods such as the Artificial Neural Network (ANN) can be used to learn and predict such compounds by training on the compounds chemical descriptors . The motivation of our work is to predict odorant molecules for the development of an artificial biosensor. In this work, we demonstrate using a set of 32 optimized odorant descriptors how an assembly of Multi-Layer Perceptrons (MLPs) can be successfully trained to differentiate among eight different chemical classes of odorant.**

In this communication, we demonstrate how it is possible to predict all 15/15 vectors from an unseen validation set with a high average prediction accuracy of 88.5% for the validation vectors. Furthermore, an introduction of a 10% noise injection level to the training set, increased the learning rate significantly as well as improve the average prediction accuracy of the MLPs to 92% for the validating vectors. Thus, this work indicates the promise of using odorant descriptor values to accurately predict chemical class and so move us forward to the realisation of an artificial odorant biosensor.

I. Introduction

"A fundamental tenet of chemistry is that the structural formula of any compound contains coded within it all that compound's chemical, physical, and biological properties" [1]. A molecular descriptor is the final result of a logic and mathematical procedure which transforms the chemical information of a molecule into a useful number, known as the descriptor index, descriptor value or simply descriptor [2]. In accordance with the similar property principle, descriptor values can be used to analyse and predict compounds and correlate structural features and chemical properties of molecules [3].

Manuscript received March 31, 2011. This work was supported by New Zealand's New Economy Research Fund C06X0701. We thank John Carlson for kindly providing raw data.

L. R. Bachtiar is a postgraduate student with the Department of Engineering Science, The University of Auckland, Auckland 1010, New Zealand. (e-mail: lbac004@aucklanduni.ac.nz).

C. P. Unsworth is a Senior Lecturer at the Department of Engineering Science, The University of Auckland, Auckland 1010, New Zealand. (e-mail: c.unsworth@auckland.ac.nz).

R. D. Newcomb is a Team Leader at The New Zealand Institute for Plant & Food Research, Private Bag 92169, Auckland 1142, New Zealand and is Associate Professor in Evolutionary Genetic at the School of Biological Sciences, The University of Auckland, Auckland 1010, New Zealand. (e-mail: Richard.Newcomb@plantandfood.co.nz).

E. J. Crampin is an Associate Professor at the Auckland Bioengineering Institute and the Department of Engineering Science, The University of Auckland, Auckland 1010, New Zealand. (e-mail: e.crampin@auckland.ac.nz).

A. Chemical Descriptors

Classical organic chemistry has long been involved with the correlation of chemical properties in terms of structure [1]. Such comparisons and correlations have been realised through the definition of chemical reference spaces which depend on the various chemical descriptors used to portray structural features and molecular properties [4]. Hundreds of descriptors that capture molecular features in various ways have been identified [5]. As depicted in Table 1, descriptors can be classified according to their dimensionality which refers to the representation of molecules from which the descriptor values were processed [4]. The performances of certain descriptors rely on their intended function and how they are applied. In turn, descriptors of different dimensionalities may be complementary in nature and not mutually exclusive [4].

EXAMPLE OF MOLECULAR DESCRIPTORS OF DIFFERENT DIMENSIONALITY	
Descriptor Index	**Definition**
1D Descriptors	
1558	Phenol / enol / carboxyl OH
1528	R--CH—X
2D Descriptors	
48	Number of benzene-like rings
22	Number of double bonds
3D Descriptors	
959	Signal 04 / weighted by atomic van der Waal volume
1321	R maximal autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities

Table I : Examples of some typical optimised descriptor values for odorants. Descriptor index values were obtained from [11] used by Haddad et all. [7].

B. Odorant Prediction Method

The motivation of our work is to predict odorant molecules for the development of an artificial biosensor. Thus, for this work a set of descriptors was used to build a predictive model for a range of chemical odorants that fell into eight specific chemical classes. Machine learning techniques such ANNs and Genetic Algorithms (GAs) and information theoretic approaches have been used for classification methods [4,6]. For this study, an Artificial Neural Networks (ANNs) approach in the form of an assembly of Multi-layer Perceptrons (MLPs) was employed to classify odorants based on their functional group, otherwise known as their chemical class. In the same conference proceedings we have investigated odorant

classification using neuron firing rates [14] rather than chemical descriptor values.

II. Odorant Classification Method

A. Descriptor Data Used

In a study conducted by Haddad et al. [7] an optimized matrix of 32 descriptors was created from a database of 1,664 descriptors. These descriptors presented the highest correlation for odorant molecules, spanning the physic-chemical space for olfaction experiments [7]. In this work, a total of 104 chemical odorants are used. They are defined by values according to the 32 descriptors. The physiochemical descriptors of each odorant molecule were obtained by generating the molecular structure for each odorant in [11] the descriptor values were then normalized [7]. Each of the 104 chemical odorants were grouped according to chemical class. A total of eight chemical classes were used described below (with the number of chemicals listed in brackets): Lactones (5 chemical odorants); Acids (15); Terpenes (16); Aldehydes (8); Ketones (6); Aromatics (13); Alcohols (17) and Esters (24).

B. Training and Implementation of an Assembly of Multi-layer Perceptrons

For this work, an assembly of MLPs was employed as the machine learning model. The MLP architecture used was that of a three-layered feed-forward network where the odorant input vector passes through successive layers: the input layer; 1 hidden layer and an output layer to produce the final network output. The input layer consisted of 32 neurons corresponding to the size of the optimized descriptors data set. The length of the hidden layer was experimentally determined to be 48 in order to obtain effective learning. A single neuron was used for the output layer to provide an output value in the range of 0 to 1. At the boundaries of each layer, input-hidden and hidden-output boundaries, a weighted sum of the neuron values was computed and passed through a binary sigmoid transfer function (f), providing outputs that approach binary limits. The weighing matrices at each boundary layer were adjusted based on the historical performance of the network, using back-propagation under supervised learning. A momentum function was included in the learning algorithm to enhance the convergence rate of the MLP [8]. A total of eight MLPs of the above architecture were implemented in Matlab software with each MLP being trained to predict a specific chemical class. Figure 1, schematically depicts how the MLP assembly is trained to predict the 8 chemical classes from the 32 optimised chemical descriptors [7]. Figure 1, highlights the MLP output for the chemical class of Esters; where the MLP is trained to give a desired output of unity when its input is that of the Ester descriptors and desired output of zero for all other chemical classes.

Due to the small number of chemicals available in each chemical class it was decided to use approximately 10% of chemicals from each chemical class as the unseen validation set. Thus, a total of 15 odorant chemicals were randomly selected for the validation set. This consisted of : Lactones (1 chemical odorant); Acids (2); Terpenes (2); Aldehydes (1); Ketones (2); Aromatics (2); Alcohols (2) and Esters (3). The remaining 89 odorants were used to train and test the eight MLPs to dedicated chemical classes. The MLPs were trained on a total of 200 shuffled epochs of the training set, towards a desired unitary output for the assigned chemical class and zero for all other classes.

C. Introducing Noise to the Prediction Model

A basic measure of an ANNs performance is its ability to generalize or properly respond to previously unseen data [10]. Poor network generalization can occur when there is insufficient training samples and higher learning parameters that the network can accommodate [9]. It has been recognized that adding noise to the learning process, commonly referred to as *noise injection*, improves the performance [13] of the ANN for a variety of situations [9,12]. A recommended optimal level [10] of ~10% noise was introduced into the training set. A total of two different training sets were used in this work: the original raw training set and a second training set with the 10% additive uniform noise.

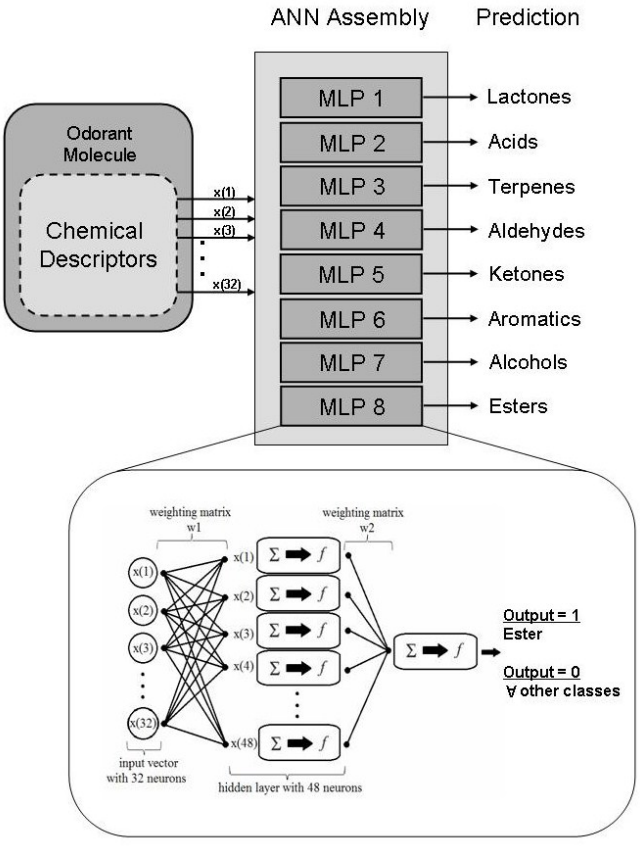

Fig. 1. Schematic of Assembly of MLPs used to predict odorant molecules. The schematic highlights the MLP output for the chemical class of Esters; where the MLP is trained to give a desired output = 1 when its input is that of the Ester descriptors and desired output = 0 for all other chemical classes.

III. RESULTS

A. Progression of Network Training and Validation Performance with Computational Time

The training error of the eight MLPs for 200 epochs is depicted in Figure 2. Similarly, the testing error of the eight MLPs is shown in Figure 3. The improvement in learning for the system with 10% added noise is clearly evident in the reduction of error in both the training and testing results of figures 2 and 3. This reinforces the strength of the noise injection as it requires less computation, (namely, less than 40 epochs to reach a performance level that supersedes the level obtained at 200 epochs using raw training data only).

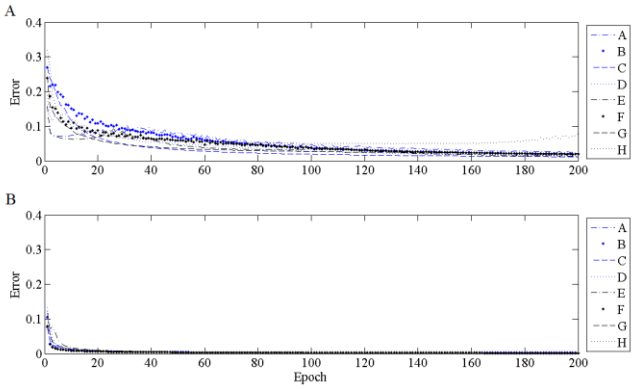

Fig. 2. Training for 200 epochs with : A) Raw and B) Noise injected data for the 8 chemical classes A-H, where: A – Lactones, B- Acids, C-Terpenes, D-Aldehydes, E-Ketones, F-Aromatics, G-Alcohols and H-Esters.

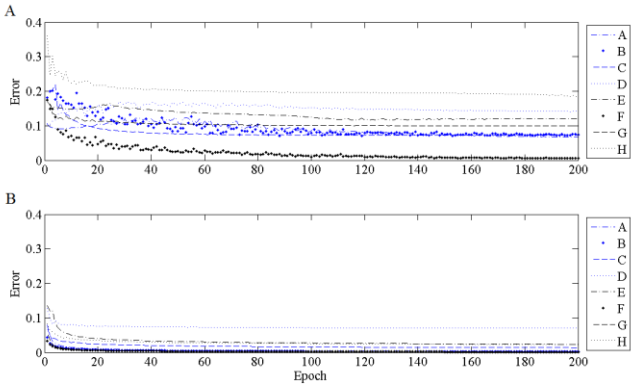

Fig. 3. Testing for 200 epochs with : A) Raw and B) Noise injected data for the 8 chemical classes A-H, where: A – Lactones, B- Acids, C-Terpenes, D-Aldehydes, E-Ketones, F-Aromatics, G-Alcohols and H-Esters.

B. Network Prediction Accuracy of Unknown Odorants

Initially a threshold level of prediction was determined. A minimum threshold level was initially determined which was the probability of randomly selecting the largest chemical class in the validation set. The largest chemical class was that of the Ester class (taking up 3 out of the 15 validation vectors). Thus, the minimum threshold level was determined to be 3/15 or 20%. For a safeguard one should chose a level comfortably over the minimum threshold. Since the results achieved were of high accuracy with a minimum prediction of 54.8% occurring for vector 9 of the aromatics group, one can choose the threshold level to be 40% prediction. Thus, avoiding the minimum threshold level of 20% and the minimum prediction of 54.5% comfortably. Table II shows the prediction accuracy of the 8 chemical classes for both raw and noise injected training sets. With the threshold level set at 40% it was possible to successfully and comfortably detect all 15/15 validation vectors using both raw and noise injected training sets with the assembly of MLPs. The performance difference of the noise injected training set over the raw training set is also given, where all but one validation vector (vector 10 of the ketones class) induces a performance increase due to the noise injected training set.

Chemical Class	Validation Vector	Prediction Accuracy (%) Raw Data	Noise Inj. Data	Improvement with Noise Inj.
Lactones	1	82.7	92.3	9.6
Acids	2	99.4	99.9	0.5
	3	99	99.7	0.7
Terpenes	4	98.8	99.5	0.7
	5	95.4	98.6	3.2
Aldehydes	6	91	98.9	7.9
Ketones	7	81.8	91	9.3
	8	63.4	75.3	11.9
Aromatics	9	54.8	55.1	0.3
	10	88.4	86.5	-1.9
Alcohols	11	98	99.6	1.6
	12	83.1	86	2.9
Esters	13	95.5	98.2	2.7
	14	98.6	99.7	1.1
	15	97.5	99.1	1.6
	Max	99.4	99.9	11.9
	Min	54.8	55.1	-1.9
	Median	95.4	98.6	1.6
	Mean	88.5	92.0	3.5
	SD	13.6	12.5	4.1

Table II : Prediction Accuracy of the 8 chemical classes for both raw and noise injected training sets. All 15/15 validation vectors were predicted with a high accuracy of prediction for both the raw and noise injected training sets. The noise injected training set clearly outperformed the raw training set in all but one class.

As can be seen in Table II, mean prediction accuracies of 88.5% for the raw and 92% for the noise injected training sets were obtained with an improvement in average prediction accuracy of 3.5% for the noise injected training set.

In addition, Table II also includes the maximum, minimum and median values of the prediction accuracy. This is to highlight the broad dynamic range, yet skewed distribution of the prediction accuracy. As can be seen most of the prediction results fall at the high end of prediction as reflected by a median of 95.4% for the raw and 98.6% for the noise injected training sets. Thus, the mean prediction accuracies of 88.5% for the raw and 92% for the noise injected training sets are conservative estimates.

IV. CONCLUSION

In this work we demonstrate how it is possible to use an assembly of Multi-Layer Perceptrons to learn and correctly

classify chemical odorants of eight different classes with high prediction accuracy. The assembly of MLPs was trained on the raw and noise injected versions of the data. Both the raw and noise injected training sets produced excellent network performance in which all validation set of 15/15 unseen odorants were successfully classified into their correct chemical class. It was found that by introducing noise injection to the training of the MLPs that a faster rate of network learning and a more accurate level of odorant identification and prediction could be achieved, producing mean prediction accuracies of 88.5% and 92% for the raw and noise injected training sets respectively. Thus, our initial work presented here shows promise for the development of artificial biosensors for the detection and correct classification of odorant molecules chemical class from their descriptor values.

REFERENCES

[1] A.R. Katritzky, M. Karelson, and V.S. Lobanov, "QSPR as a means of predicting and understanding chemical and physical properties in terms of structure", *Pure & Applied Chemistry*, vol. 69, pp. 245–248, 1997.

[2] M.P. González, C. Terán, M. Teijeira, and P. Besada, "Geometry, topology, and atom-weights assembly descriptors to predicting A1 adenosine receptors agonists", *Bioorganic & medicinal chemistry letters*, vol. 15, pp. 2641-5, May 2005

[3] J. Bajorath, "Selected Concepts and Investigations in Compound Classification , Molecular Descriptor Analysis and Virtual Screening," *Society*, pp. 233-245, 2001.

[4] J.W. Godden, J.R. Furr, L. Xue, F.L. Stahura, "Molecular Similarity Analysis and Virtual Screening by Mapping of Consensus Positions in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality," *Journal of Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 21–29, 2004.

[5] L. Xue and J. Bajorath, "Molecular Descriptors in Chemoinformatics , Computational Combinatorial Chemistry , and Virtual Screening", *Combinatorial Chemistry & High Throughput Screening*, pp. 363-372, 2000.

[6] J. Gasteiger and J. Zupan, "Neural Networks in Chemistry," *Angewandte Chemie International Edition in English*, vol. 32, pp. 503-527, Apr. 1993.

[7] R. Haddad, R. Khan, Y.K. Takahashi, K. Mori, D. Harel, and N. Sobel, "A metric for odorant comparison.", *Nature methods*, vol. 5, pp. 425-9, May 2008.

[8] S. Huang, K.K. Tan, and K.Z. Tang,"Neural network control: theory and applications", *Research Studies Press,* 2004.

[9] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training", *IEEE transactions on neural networks - a publication of the IEEE Neural Networks Council*, vol. 3, pp. 24-38, Jan. 1992.

[10] R.I. Levin, N.A.J. Lieven, M.H. Lowenberg, "Measuring and Improving Neural Network Generalization for Model Updating," *Journal of Sound and Vibration*, vol. 238, pp. 401-424, Nov. 2000.

[11] PubChem (httpe://pubchem.ncbi.nlm.nih.gov/search/) and entering it into Dragon (http://www.talete.mi.it/download.htm) software.

[12] C.P. Unsworth, G.G. Coghill, "Excessive noise injection training of neural networks for markerless tracking in obscured and segmented environments",*Neural Computation*, vol.18, no.9, pp. 2122-45, 2006.

[13] J. Sietsma, R.J.F. Dow,"Creating artificial neural networks that generalize", *Neural Networks,* vol 4, pp.67–79, 1991.

[14] L.R. Bachtiar, C.P. Unsworth, R.D. Newcomb, E.J. Crampin, "Using artificial neural networks to classify unknown volatile chemicals from the firings of insect olfactory sensory neurons", *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Accepted).