Structure-musk odour relationship studies of tetralin and indan compounds using neural networks

Driss Cherqaoui,[a] M'Hamed Esseffar,[a] Didier Villemin,[b] Jean-Michel Cense,[c] Maurice Chastrette,[d] and Driss Zakarya*,[e]

[a] *Département de Chimie, Faculté des Sciences Semlalia, BP S 15, Université Cadi Ayyad, Marrakech, Morocco.*
[b] *Ecole Nationale Supérieure d'Ingénieurs de Caen (ENSI de Caen) ISMRA (CNRS URA 480), 6 boulevard Maréchal Juin, 14050 Caen cédex, France.*
[c] *Ecole Nationale Supérieure de Chimie de Paris, 11 rue P. et M. Curie, 75005 Paris, France.*
[d] *Laboratoire de Chimie Organique Physique et Synthétique (CNRS URA 463), Université Claude-Bernard-Lyon I, 69622 Villeurbanne cédex, France.*
[e] *Département de Chimie, Faculté des Sciences et Techniques, B.P. 146, Mohammadia, Morocco.*

Models of the relationships between structure and musk odour of tetralin and indan compounds were elaborated with a multilayer neural network using the back-propagation algorithm. The neural network was used to classify the compounds studied into two categories (musk or non-musk). The cross-validation procedure was used to assess the predictive power of the network. Each molecule was described by eight global parameters: five steric and three electronic descriptors. The neural network's results were successfully compared to those given by the *k*-Nearest Neighbours and the Bayesean methods, both in the classification and prediction tests. The contribution of each descriptor to the structure-odour relationships was evaluated. Three out of the eight descriptors were thus found to be the most relevant in the molecular description for the prediction of musk odour. This research points out that neural networks are likely to become a useful technique for structure-odour relationships.

Musk was extracted from animals for many decades but using natural products on an industrial scale proved limited and researchers thus started to get interested in synthetic musk. The first synthetic musk was obtained by Bauer.[1] Since then, the impressive number of theoretical and experimental studies devoted to the making of new synthetic musk products has shown their growing importance in perfumery.

The study of relationships between molecular structures and odour has hastened the development of new musk compounds. A certain number of computational techniques have been found useful for the establishment of these relationships. A relatively recent technique and one that shows considerable promise is that of Neural Networks (NNs).[2,3] NNs have arisen from the fascination of mankind for the understanding and emulating the human brain's ability to solve various problems such as pattern and character recognition. There has been great interest in applying the NN approach to chemical problems.[4–9] However, NN applications to structure-odour relationships are quite sparse.[10–14] More recently, structure-musk odour relationships[11,12] have been established by means of NNs on a sample of carbonyl tetralins and indans. Each molecule was described by steric and electronic descriptors, *coding substituents on a common skeleton*. It was shown that it is possible to model the relationships between the molecular structure and the musk odour of the compounds studied. However, coding molecular structure from the substituents only can raise certain problems in the establishment of Structure-Activity Relationships (SAR): (*i*) As the number of substituents increases, so does the number of descriptors. As a consequence, a lot of data are needed to model the SAR, which cannot always be obtained. (*ii*) The common skeleton's role is not taken into account in the SAR. The results obtained may thus not be satisfactory. (*iii*) Some descriptors' values of uncommon substituents cannot always be found in the literature. The size of the database is therefore reduced.

The goal of the current work is: (*i*) to provide an application of NNs to the structure-musk odour relationships of a set of 85 tetralins and indans, using *global molecular properties*, which were revealed to be important in the recognition of odorants by olfactory systems,[10] (*ii*) to call attention to the interest of these properties for the molecular structure description, (*iii*) to compare the results obtained by the NNs to those of the *k*-Nearest Neighbours (*k*NN) and Bayes methods, and (*iv*) to measure the contribution of each descriptor to structure-musk odour relationships.

Methods

Compounds studied

A series of 32 indans and 53 tetralins (see Fig. 1 and Table 1) were taken under consideration. The set of tetralins consisted of 35 musks and 18 non-musks and that of indans consisted of 16 musks and 16 non-musks. The structures and olfactory properties of these compounds were described previously[10,11,15,16] and are listed in Table 1. The odour is coded M if the molecule is musk and NM if otherwise. The information about odour intensities was not precise enough to be taken into account.

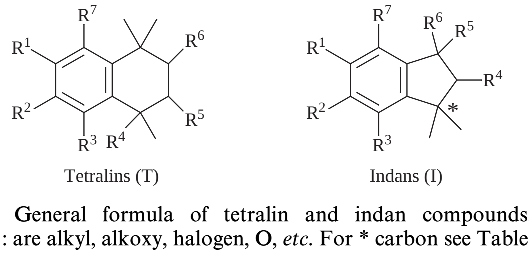

Fig. 1 General formula of tetralin and indan compounds. R_1, R_2–R_7: are alkyl, alkoxy, halogen, O, *etc.* For * carbon see Table 1

Table 1 Structures and musk odour of indan (I) and tetralin (T) compounds (M for musk compounds and NM for non-musk compounds)

Compound	R^1	R^2	R^3	R^4	R^5	R^6	R^7	Odour
I1	CHO	Me	H	Me	Me	Me	H	M
I2	Ac	Me	H	Me	Me	Me	H	M
I3	Ac	Et	H	Me	Me	Me	H	M
I4	C(O)Et	Me	H	Me	Me	Me	H	M
I5	Ac	Et	H	H	Me	Me	H	M
I6	C(O)Et	Me	H	H	Me	Me	H	M
I7	Ac	H	H	H	Me	Me	H	NM
I8	Ac	Pr^i	H	H	Me	Me	H	NM
I9	Ac	Et	H	Et	Me	Me	H	M
I10	Ac	Me	H	H	Me	Me	H	M
I11	Ac	Me	H	Et	Me	Me	H	M
I12	Ac	H	H	Me	Me	Me	H	M
I13	Ac	Me	H	Me	Me	H	H	M
I14	Me	Me	H	Me	Me	Me	Ac	M
I15	CN	Me	H	Me	Me	Me	H	M
I16	Ac	Me	H	Me	Pr^i	H	H	M
I17	CHO	Me	H	Me	Pr^i	H	H	M
I18	Ac	Me	H	Et	Pr^i	H	H	M
I19	Ac	Et	H	Et	Pr^i	H	H	NM
I20	Me	Ac	H	H	Pr	Me	H	NM
I21	Me	Ac	H	H	Bu^i	Me	H	NM
I22	Me	Ac	H	H	Hex^i	Me	H	NM
I23[a]	Bt^t	H	H	H	O	—	H	NM
I24[a]	Ac	H	Bt^t	H	H	H	H	NM
I25	H	Bt^t	H	H	O	—	H	NM
I26	H	H	H	H	H	H	Ac	NM
I27	Ac	H	H	H	H	H	H	NM
I28	H	Ac	H	H	H	H	H	NM
I29	H	Bt^t	H	H	H	H	CO$_2$H	NM
I30	H	H	H	H	H	H	CO$_2$H	NM
I31	Me	Ac	H	H	Me	H	H	NM
I32[b]	Bt^t	H	Et	—	Me	Me	H	NM
T1	CHO	H	H	Me	H	H	H	NM
T2	CHO	Me	H	Me	H	H	H	M
T3	CHO	Et	H	Me	H	H	H	M
T4	CHO	Pr^i	H	Me	H	H	H	M
T5	CHO	OMe	H	Me	H	H	H	NM
T6	CHO	H	OMe	Me	H	H	H	NM
T7	CHO	Me	H	Me	H	H	Me	M
T8	CHO	H	Me	Me	H	H	Me	NM
T9	CHO	Me	Me	Me	H	H	H	M
T10	CHO	Me	H	Me	H	H	OMe	M
T11	CHO	Et	H	Me	H	H	OMe	M
T12	CHO	H	Me	Me	H	H	OMe	M
T13	CHO	Me	Mecy	Me	H	H	H	M
T14	CHO	Me	Me	Me	H	H	Me	NM
T15	CHO	Me	Me	Me	H	H	OMe	M
T16	CHO	Me	H	Me	Me	H	H	M
T17	CHO	Me	H	Me	Me	H	Me	M
T18	CHO	Me	H	Me	H	Me	Me	M
T19	CHO	Me	H	Me	Me	Me	H	M
T20	CHO	Me	H	Me	Me	Me	H	M
T21	CHO	Me	H	Me	Mecy	H	H	M
T22	Ac	Me	H	Me	H	H	H	M
T23	Ac	Et	H	Me	H	H	H	M
T24	Ac	Pr^i	H	Me	H	H	H	NM
T25	Ac	OMe	H	Me	H	H	H	NM
T26	Ac	F	H	Me	H	H	H	NM
T27	Ac	Cl	H	Me	H	H	H	NM
T28	Ac	CN	H	Me	H	H	H	NM
T30	Ac	CO$_2$Me	H	Me	H	H	H	NM
T31	Ac	CH$_2$OMe	H	Me	H	H	H	NM
T32	Ac	OH	H	Me	H	H	H	NM
T33	C(O)Et	Me	H	Me	H	H	H	M
T34	C(O)Et	Et	H	Me	H	H	H	M
T35	C(O)Et	Pr^i	H	Me	H	H	H	NM
T36	C(O)Pr^i	Et	H	Me	H	H	H	NM
T37	Ac	H	H	Me	Me	H	H	M
T38	Ac	Me	H	Me	Me	H	H	M
T39	Ac	Me	H	Me	Et	H	H	M
T40	Ac	Me	Mecy	Me	H	H	H	M
T41	Ac	Pr^i	H	Me	Me	H	H	NM
T42	Ac	Me	H	Me	Mecy	H	H	M
T43	Ac	Me	H	Me	Me	Me	H	M
T44	Ac	Et	H	Me	Me	H	H	M

Table 1 (*Continued*)

Compound	R^1	R^2	R^3	R^4	R^5	R^6	R^7	Odour
T45	CHO	Me	H	Me	H	H	OH	M
T46	CN	Me	H	Me	H	H	OMe	M
T47	CN	Et	H	Me	H	H	OMe	M
T48	CN	Me	H	Me	H	H	H	M
T49	CN	Et	H	Me	H	H	H	M
T50	Ac	Me	H	Et	H	H	H	M
T51	H	H	H	Me	O	H	H	NM
T52[c]	Bt^{t}	H	Ac	H	H	H	H	M
T53[c]	Pr^{i}	Me	Ac	H	H	H	H	M

Me, Et, Pr, Pr^{i}, Hex^{i}, Bt^{t}, Ac, Mecy and O stand for methyl, ethyl, *n*-propyl, isopropyl, isobutyl, isohexyl, *tert*-butyl, acetyl, cyclic —CH_2—90 group and =O. [a] The carbon * (see Fig. 1) is substituted by two H. [b] Double bond between the carbon * and the carbon formerly substituted by R^4; the carbon * is only substituted by H. [c] The carbon substituted by R^4 is also substituted by H instead of Me.

Molecular descriptors used

Structure-musk odour relationships have already been studied using both empirical and theoretical methods. Several papers[11,12,17–20] have thus shown that the musk odour might be closely related to the size, shape, functionality and/or hydrophobicity of a molecule. These four factors can be described by the molecular surface (srf), the molecular volume (vlm), the ovality (o) of the molecule, the molecular weight (mw), the number of carbon atoms (nc), the dipole moment (dp), the square root of the mean of squared charges on oxygen atoms (qo) and the highest occupied molecular orbital (HOMO). We based this choice on the conclusions of recent studies.[14,21,22] Zakarya *et al.*[14] stressed the importance of these parameters for the prediction of sandalwood odour. Cense *et al.*[21] and Bodor and Huang[22] showed that the compounds' hydrophobic properties depend on srf, o, dp, qo, nc and mw.

The geometries of the molecules were first designed using HyperChem.[23] They were then fully optimised using the AM1 Hamiltonian[24] with AMPAC on a Silicon Graphics workstation.

Geometric parameters were computed by numerical integration techniques similar to those described by Gavezzotti.[25] To determine the volume, the molecule was immersed in a regular three-dimensional grid bounded by the three-dimensional extent of the molecule and every grid point was tested for inclusion into any atom of the molecule. The grid spacing was 0.1 Å. The molecular surface was calculated from a regular grid of points disposed on the surface of each atom: 4,558 points were used for hydrogen atoms and 10,270 points were used for other atoms. The ovality estimation was that given by Bodor:[26]

$$\mathrm{o} = \mathrm{srf}/(4\pi K) \tag{1a}$$

$$K = (3\mathrm{vlm}/4\pi)^{2/3} \tag{1b}$$

The dipole moment, charge, molecular weight, HOMO and number of carbon atoms were directly read from the AMPAC output.

Neural network (NN)

The NN (see Fig. 2) trained by the back-propagation (BP) of errors[3] had the following architecture: the input layer consisted of nine neurons, eight of which represented the eight molecular descriptors and the last neuron represented a bias. The output layer consisted of one neuron, namely the olfactorial activity. The NN had a feedforward layered structure with connections allowed only between adjacent layers. The musk odour represented by the output of the NN was coded 0.9 if the molecule was musk and 0.1 otherwise. Input data were normalised between 0.1 and 0.9.

The sigmoidal transfer function used in the NN is given by

$$O_i = [1 + \exp(-\Sigma W_{ij} O_j)]^{-1} \tag{2}$$

where O_i and O_j are the outputs of neuron *i* and *j*, respectively, and W_{ij} is the weight connecting neuron *i* to neuron *j*.

The weights of connections between the neurons were initially assigned with random values uniformly distributed between -0.5 and $+0.5$ and no momentum was added. The BP algorithm was used to adjust these weights. The learning rate was initially set to 1 and was gradually decreased until the error function could no longer be minimised.

k-Nearest Neighbours (*k*NN)

The *k*NN method[27] is one of the most widely used classification techniques. In this method, the distance (usually Euclidean) between the pattern vector of an unknown compound and each of the pattern vectors of the training set is first computed. The *k* nearest neighbours to the unknown compound are then selected and the latter is classified as belonging to the class of the majority of the *k* nearest neighbours in the training set. In this paper, we used $k = 1$ and the Euclidean distance between pattern vectors.

Bayes method

In the Bayesian approach,[27] the decision rule assigns the observation vector x ($x \in \mathrm{R}^n$) to the class having the highest

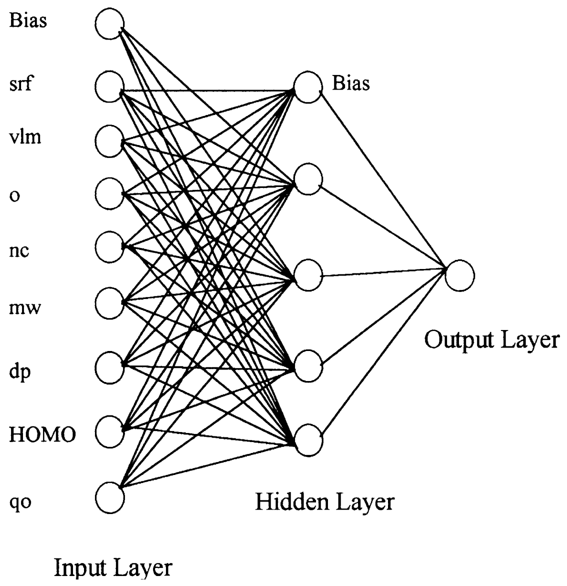

Fig. 2 Three-layer Neural Network. The architecture shown is 8-4-1. The bias unit is not considered in the unit count

a posteriori probability $p(C_i/x)$ among all the classes C_i:

$$p(C_i/x) = p(C_i) \cdot f(x/C_i)$$

$p(C_i)$ are *a priori* probabilities and are known or fixed in advance and $f(x/C_i)$ is the conditional density function. It is assigned to each class and is assumed to be Gaussian:

$$f(x/C_i) = \frac{\exp[-(1/2)(x - \mu_i)^t \sum_i^{-1} (x - \mu_i)]}{(2\pi)^{n/2} |\sum_i|^{1/2}} \quad (4)$$

μ_i and $\sum_i$ are respectively the mean vector and the covariance matrix of the elements of each class C_i in the learning set.

In this paper, $p(C_{\text{musk}}) = p(C_{\text{non-musk}}) = 1/2$.

All calculations using the NN, *k*NN and Bayes methods were carried out on an 166 MHz Pentium computer using our programs written in C language.

Results and Discussion

Three studies were achieved: classification, estimation of the prediction error and the descriptors' contribution. Classification consists in using a trained NN to recognise the odour of molecules present in the training data set, whereas estimation of the prediction error corresponds to the use of an NN to predict the odour of molecules not included in the training set. The third objective of this work is an evaluation of the importance of molecular descriptors in the establishment of structure-musk odour relationships.

Classification

It is well-known that the performance of an NN depends on many variables, including the molecular descriptors, the number of hidden neurons and the degree of homology between the training and the test sets. In order to determine the optimum number of hidden neurons and to make sure that the eight descriptors retained were adequate, several training sessions were carried out with different numbers of neurons in the hidden layer. The configuration of the six NNs was 8-*x*-1, where *x* ($x = 4$–9) represented the number of hidden neurons. After the training phase, the classification ability of the NN was applied to the same 85 compounds. All compounds were correctly recognised in less than 1,000 iterations with the six architectures tested. The Bayes method was applied to the same data set and the same eight molecular descriptors and gave a classification rate of 91.8%. This preliminary study enabled us to conclude that all the NN architectures were able to establish a satisfactory relationship between the molecular descriptors and the musk odour.

Estimation of the prediction error

One of the most important attributes of NNs is their ability to generalise, that is their ability to make reliable predictions on new data with an accuracy similar to that obtained with the training set. After determining the range of hidden neurons giving a good classification, we turned to the most important aspect of NNs: their predictive ability. The cross-validation method was used to estimate which NN architecture would be best suited for a given data set of 85 compounds, or which NN would result in a classifier having the lowest prediction error. Each 8-*x*-1 network was trained *N* times. After each training with $N - 1$ objects, always leaving out a different object, a different classifier was obtained and this classifier was tested with the remaining object, whether the classification was correct or not. The NN architecture (8-6-1 in this case) yielding the smallest estimated prediction errors was selected for building the final classifier model. The cross-validation method was also used in the case of the *k*NN and Bayes methods.

Among all the NN architectures, the best one is 8-6-1. The estimation of the prediction error of the 8-6-1 model is 91.8%. The incorrectly classified molecules correspond to molecules I32, T6, T8, T14, T24 and T25 for the non-musk compounds and to molecule T45 for the musk compounds. The results are satisfying and show once again that the eight molecular descriptors are particularly relevant. The NN has a superior performance to those of the *k*NN ($k = 1$, 88.2%) and Bayes methods (89.4%). This does not mean that the NN is a polynomial model but that it is able to extract information from examples in order to develop an internal representation of the musk odour without explicitly incorporating rules into the network. As to the *k*NN method, increasing the number of neighbours (*k*) does not improve the results.

Contribution of descriptors to the classification ability

The evaluation of the relevance of the descriptors proved quite interesting and useful. This is why we chose to estimate their relative contribution in two different ways.

(*i*) The contribution of descriptor *i* ($i = 1$–8) was estimated from the trained 8-6-1 configuration network. The descriptor under study was removed from the 8-6-1 NN together with its corresponding weights. Then the network (7-6-1) calculated the output of each molecule as usual. The mean of the deviations' absolute values Δm_i between the observed musk odour (0 or 1) and the estimated odour for all compounds was calculated. This process was reiterated for each descriptor. Finally, the contribution (C_i) of descriptor *i* is given by:

$$C_i = 100 \cdot \Delta m_i \Big/ \sum_{i=1}^{8} \Delta m_i \quad (5)$$

(*ii*) The contribution of each descriptor to the classification ability was measured thanks to Chastrette *et al.*'s third method.[11]

Table 2 shows that these two methods give the same results. These results also indicate that the relative importance of the descriptors varied in the following order: nc > vlm > dp > mw > qo = srf > o > HOMO. In order to assess their importance, the first three descriptors (nc, vlm and dp) were used in a 3-9-1 NN. Nine hidden neurons were chosen so as to maintain ρ [ρ = (number of data points in the training set)/(number of adjustable weights controlled by the network)] between 1 and 2.2, as recommended by Andrea and Kalayeh.[28] The cross-validation results obtained (86% with 800 epochs) show the relevance of these three descriptors.

Conclusion

We have shown the applicability of the NN approach to musk odour classification of tetralin and indan compounds, using *global molecular properties.* The performances of the NNs were compared with those of the *k*NN and Bayes methods and proved to be better. The contribution of the descriptors to

Table 2 Contributons (C_i) of steric and electronic descriptors to the classification

Descriptor	Ci[a] (%)	Ci[b] (%)
nc	20.8	20.6
vlm	19.3	19.4
dp	17.8	18.5
mw	11.4	11.3
qo	8.6	8.5
srf	8.6	8.5
o	8.1	8.1
HOMO	5.5	5.2

[a] Given by the first method [eq. (5)] described in the text. [b] Given by the second method[11] described in the text.

structure-musk odour relationships was evaluated, leading to a classification of the descriptors used according to their importance. This provides supplementary information to understand olfaction mechanisms.

The results obtained confirm the findings of the previous study according to which musk odour is related to steric, electronic and hydrophobic effects. The NN approach would seem to have great potential for determining structure–odour relationships.

References

1 A. Bauer, *Ber. Dtsch. Chem. Ges.*, 1891, **24**, 2832.

2 D. E. Rumelhart and C. E. Hinton, *Nature* (*London*), 1986, **323**, 533.

3 J. A. Freeman and D. M. Skapura, *Neural Networks: Algorithms, Applications, and Programming Techniques*, Addison-Wesley Publishing, Reading, MA, 1991, pp. 89–132.

4 J. A. Burns and G. M. Whitesides, *Chem. Rev.*, 1993, **93**, 2583.

5 M. Tusar, J. Zupan and J. Gasteiger, *J. Chim. Phys.*, 1992, **89**, 1517.

6 D. Villemin, D. Cherqaoui and A. Mesbah, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1288.

7 O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye and J. P. Doucet, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 644.

8 M. Chastrette, J. Y. De Saint Laumer and J. F. Peyraud, *SAR QSAR Environ. Res.*, 1992, **1**, 221.

9 D. Zakarya, L. Farhaoui and S. Fkih-Tétouani, *J. Phys. Org. Chem.*, 1996, **9**, 672.

10 D. Zakarya, Ph.D. Thesis, Université de Lyon I, 1988.

11 M. Chastrette, D. Zakarya and J. F. Peyraud, *Eur. J. Med. Chem.*, 1994, **29**, 343.

12 M. Chastrette, C. E. Aïdi and J. F. Peyraud, *Eur. J. Med. Chem.*, 1995, **30**, 679.

13 M. Chastrette, D. Cretin and C. E. Aïdi, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 108 and references cited therein.

14 D. Zakarya, D. Cherqaoui, M. Esseffar, D. Villemin and J. M. Cense, *J. Phys. Org. Chem.*, 1997, **10**, 612.

15 B. Bersuker, A. S. Dimoglo, M. Y. Gorbachov, P. F. Vlad and M. Pesaro, *New J. Chem.*, 1991, **15**, 307.

16 C. Fehr, J. Galindo, R. Haubrichs and R. Perret, *Helv. Chim. Acta*, 1989, **72**, 1537.

17 J. E. Amoore, *Perf. Ess. Oil Rec.*, 1952, **43**, 321.

18 E. T. Theimer and J. T. Davis, *J. Agric. Food Chem.*, 1967, **15**, 6.

19 M. G. J. Beets, *Structure-activity Relationships in Human Chemoreception*, Applied Science Publishers, London, 1978, pp. 407–425.

20 G. Klopman and D. Ptchelintsev, *J. Agric. Food Chem.*, 1992, **40**, 2244.

21 J. M. Cense, B. Diawara, J. J. Legendre and G. Roullet, *Chem. Intell. Labor. Sys.*, 1994, **23**, 301.

22 N. Bodor and M. J. Huang, *J. Am. Chem. Soc.*, 1989, **111**, 3783.

23 *HyperChem*, Hypercube, Inc, Waterloo, Ontario, Canada.

24 M. J. S. Dewar, E. G. Zoebisc, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.

25 A. Gavezzotti, *J. Am. Chem. Soc.*, 1983, **105**, 5220.

26 N. Bodor, A. Harget and M. J. Huang, *J. Am. Chem. Soc.*, 1991, **113**, 9480.

27 D. Michie, D. J. Spiegelhalter and C. C. Taylor, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Publshers, New York, 1994.

28 T. A. Andrea and H. Kalayeh, *J. Med. Chem.*, 1991, **34**, 2824.

Received in Orsay, France, 28th May 1997; *Revised manuscript received* 10*th November* 1997; *Paper* 7/09269E