

Mutagenicity of aminoazo dyes and their reductive-cleavage metabolites: a QSAR/QPAR investigation

Les Sztandera, Ashish Garg, Seth Hayik,
Krishna L. Bhat, Charles W. Bock*

*Department of Chemistry & Biochemistry,
School of Science and Health and Department of Computer Science and Information Systems,
Philadelphia University, School House Lane and Henry Avenue, Philadelphia, PA 19144 USA*

Received 29 November 2002; accepted 9 May 2003

Abstract

Quantitative structure–activity/property–activity relationships are developed that correlate the observed mutagenic behavior of 62 aminoazo derivatives and 12 of their reductive cleavage products with a variety of molecular descriptors calculated using quantum-chemical semiempirical methodology. Multilinear regression techniques using 8 descriptors are shown to account for more than 70% of the variation in the relative mutagenic activity of these compounds. Approaches using artificial neural networks in conjunction with fuzzy logic can account for about 95% of this variation using 8 descriptors.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Aminoazo dyes; Reductive-cleavage products; Quantitative structure–activity relationships; Mutagenicity; Artificial neural networks

1. Introduction

It is well known that some azo dyes are both mutagenic and carcinogenic e.g., 6-dimethylamino-phenylazobenzothiazole (6BT) is a strong mutagen in the *Salmonella typhimurium* TA98 bacterial tester strain in the presence of an induced rat-liver S9 mix (TA98 + S9), and a potent liver carcinogen in rodents [1]. In stark contrast, other azo dyes are neither mutagenic nor carcinogenic under quite similar conditions, e.g., 4'-phenyl-4-

dimethylaminoazobenzene (4'-Ph-DAB) [2]. There are also some azo dyes that have been shown to be mutagenic but not carcinogenic, e.g., 5-dimethylaminophenyl-azoindoline (51N) is a strong mutagen in TA98 + S9, however, it does not appear to be a rodent carcinogen [1].

Several mechanisms for the carcinogenicity of azo dyes have been identified in the literature [1]. These mechanisms, which may be compound specific, generally incorporate some form of metabolic activation of the dye to reactive electrophilic intermediates that covalently bind to biological macromolecules. For example, azo compounds that also contain amino subgroups, such as 4-aminoazobenzene (AAB) or *N*-methyl-4-aminoazobenzene (MAB), may

* Corresponding author. Tel.: +1-215-951-2876; fax: +1-215-951-6812.

E-mail address: bockc@philau.edu (C.W. Bock).

become activated by metabolic oxidation of this particular subgroup. We recently developed [3] several quantitative structure–activity/property–activity relationships (QSAR/QPARs) for the observed mutagenic activity in TA98+S9 for a collection of 34 aminoazobenzene dyes and 9 of their potential *N*-hydroxy and ester metabolites. The aminoazo compounds in this preliminary investigation were limited to those that contained a single azo linkage and excluded dyes with solubilizing groups, such as sulfonic acid. These restrictions were imposed to reduce the structural diversity of compounds in the study and, thus, increase the likelihood of developing satisfactory mutagenicity correlations with only a few molecular descriptors. Models based on multilinear regression and neural network techniques were able to account for as much as 80% of the variation in the reported mutagenic behavior of the 43 aminoazobenzene compounds in the data set, using only 3–5 descriptors.

Although there is ample evidence that many aminoazobenzene dyes are inherently toxic in their pure form, they are also susceptible to reductive cleavage of the azo linkage, primarily by enzymes from intestinal bacteria. Most of the compounds released by this reductive cleavage mechanism are substituted anilines, which are then exposed to possible activation by *N*-hydroxylation of their amino subgroup(s); many substituted anilines are well known to be mutagenic and/or carcinogenic [4–6]. It should be noted that the extent to which *Salmonella* tester strains directly reduce azo dyes to their free amines is not entirely clear [1,7,8].

The purpose of this paper is to derive QPARs for the observed mutagenicity in TA98+S9 of an enhanced collection of azo dyes, as well as, their reductive cleavage products; *N*-hydroxy and ester metabolites of these compounds have also been included where data are available. All the compounds in this enhanced data set still contain at least one amino subgroup, however, dyes that contain a sulfonic acid group [9] and dyes that have two azo linkages [10] have also been included if quantitative mutagenicity data was available. The inclusion of such structural diversity presents a significant challenge for QSAR/QPAR development.

2. Methods

We employed the program CODESSA [11], in conjunction with the program AMPAC 5.0, [12] to develop correlation equations for 47 aminoazobenzene derivatives, 15 substituted disazo dyes that also contain a free amine group, and 12 of their reductive-cleavage products. The CODESSA software system has been used successfully to establish correlations in a wide variety of applications [13–16]. The structures of all the compounds involved in this study were fully optimized at the semiempirical AM1 computational level as implemented in AMPAC 5.0; many of these compounds have numerous local minima on their potential energy surfaces and extensive searches were performed to locate those conformers with the lowest heat of formation, ΔH_f . We have previously reported [17–19] some structural and electronic properties of several of the aminoazobenzene derivatives in this study, calculated using density functional theory (DFT) [20].

The CODESSA/AMPAC integrated software package [11] was used to calculate hundreds of molecular descriptors (constitutional, topological, geometrical, electrostatic, quantum-chemical, and thermodynamic) for each of the compounds in this study at their AM1 optimized geometries. The log of the octanol-water partition coefficient, $\text{Log}P$ [21], which has been implicated in a variety of mutagenicity QSAR/QPARs [22,23], and the log of the aqueous solubility, $\text{Log}S$, were calculated using the additive-constitutive approach implemented in the $\text{Log}D$ suite (version 4.5) from Advanced Chemical Development, Inc. [24], and added to the pool of descriptors. The entire collection of descriptors was then used in conjunction with the statistical facilities of CODESSA to develop multilinear regression models for the log of the measured mutagenicity (rev/nmol) in TA98+S9, LogTA98 .

Artificial neural networks (ANNs) are rapidly becoming the method of choice for QSAR/QPAR development. ANNs are model-free mapping devices capable of capturing complex nonlinear relationships in data that may be missed by conventional multilinear regression techniques. The particular approach we adopted in this investigation

integrates fuzzy logic with ANNs [25–27]. These two methodologies effectively complement one other: neural networks supply the computational power necessary to process rapidly large quantities of data, while fuzzy logic provides a high-level reasoning capability that guides the overall construction of the network. The algorithm we employed generates a feed-forward network architecture for a given data set and, after generating fuzzy entropies at each node of the network, it switches to fuzzy decision making based on those entropies. Nodes and hidden layers are added as needed until the learning task is accomplished; in this study we restricted the architecture to a single hidden layer. A more complete description of the FCID3 algorithm we used can be found in reference [28]. This approach to developing neural networks represents a significant improvement over that employed in our previous study [3].

3. Results and discussion

Observed values of the mutagenicity in the TA98 + S9 system for the aminoazobenzene derivatives, disazo derivatives, and reductive cleavage products included in this study are listed in Tables 1–3 respectively [9,10,29–43]; values of $\text{Log}P$, $\text{Log}S$, and melting points (where available) are also included in these Tables. The quantitative mutagenicity data we have used come from a variety of laboratories, over an extended period of time and are likely to involve some “noise,” e.g. the mutagenicity of 3'-Me-MAB has been reported as 445 rev/nmol [33] and also as 233 rev/nmol [38]. Fortunately, the mutagenic activity of the 74 compounds in these tables range over some 5 orders of magnitude, from about 10^{-3} to 10^2 rev/nmol, and thus, provide a broad range of input values for QSAR/QPAR development.

A number of the dyes we used in this study are typically employed as sodium or potassium salts. In these cases, we optimized their anionic forms to avoid difficulties in determining the precise location of the Na^+ or K^+ counterions. Thus, for these dyes the values of the descriptors calculated by CODESSA and used for our correlation studies

refer to their ionic forms. To obtain a value of $\text{Log}P$ for these salts, however, we used the free acid structure in the LogD software [24].

The calculated values of $\text{Log}P$ (at pH = 7) for the molecules in Tables 1–3 range from –2.6 to 5.2, showing that these compounds have a broad spectrum of hydrophilic/hydrophobic character. For comparison, we note that values of $\text{Log}P$ for most drugs developed by the pharmaceutical industry are in the range 0–5 [44]. Values of $\text{Log}S$ (at pH = 7) for the molecules in Table 1–3, were initially calculated without using experimental melting point data. When values of the melting points of these azo derivatives could be found in the literature, their aqueous solubilities were recalculated using this additional data, and the resulting, generally more reliable [24], values of $\text{Log}S$ are also listed in Tables 1–3.

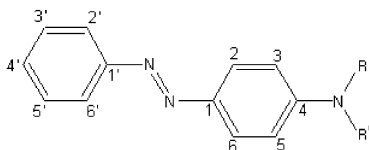
In cases where the melting point is unusually high, the predicted solubility is generally lower than that found in the absence of melting point data, e.g. the melting point of 3'-COOH-MAB has been reported as 207–208 °C and, using this data, the calculated value of $\text{Log}S$ is reduced from 1.00 to 0.16. The predicted values of $\text{Log}S$ in Tables 1–3 clearly show the greater aqueous solubility of the reductive-cleavage products compared to the parent dyes. Furthermore, they serve to quantify the enhanced solubility imparted to dyes by the presence of a sulfonic acid group.

4. Multilinear regression equations

The statistical models we developed primarily employed the best multilinear regression (BMLR) method implemented in CODESSA [11]; this method selects the best two-descriptor regression model, the best three-descriptor regression model, etc., based on the highest value of the square of the regression coefficient, R^2 . The models are constituted from a reduced set of non-collinear descriptors as determined by the pair correlation matrix. The heuristic method implemented in CODESSA [11] was used in a few instances to provide alternative correlation equations with particular descriptors forced into the model, as well as, to increase the number and the diversity of

Table 1

Observed mutagenicity (rev/nmol) in the TA98 *Salmonella typhimurium* bacterial strain with S9 activation, calculated values [24] of Log*P*, Log*S* and melting points (°C), for derivatives of (A) AAB, (B) MAB, (C) DAB, (D) their metabolites, and (E) their sulfonic acid salts



Compounds	Mutagenic activity in TA98 + S9 (rev/nmol) [Ref.]	Log <i>P</i> ^a	Log <i>S</i> ^b	Melting point (°C) [Ref]
<i>A. AAB (R=R'=H)</i>				
4'-NEt ₂ -3-OMe-AAB	0.007 [10]	5.16	-3.41(-3.77)	147–149 [29]
2-OMe-AAB	0.010 [30]	3.87	-1.85(-2.43)	157–159 [31]
4'-OH-AAB	0.053 [9]	2.55	-0.66(-1.34)	180–181 [32]
3'-Me-4'-OH-AAB	0.059 [33]	3.01	-1.14	^c
4'-OH-2',3-diMe-AAB (4'-OH-OAT)	0.112 [34]	3.47	-1.62	^c
AAB	0.204 [10]	3.13	-1.05(-1.39)	124–125 [35]
3'-Me-AAB	0.240 [33]	3.59	-1.52(-1.57)	89–91 [32]
3-OMe-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	0.390 [10]	2.58	-1.48(-1.32)	129–131 [35]
3'-CH ₂ OH-AAB	0.596 [33]	1.94	-0.25(-0.23)	107–109 [36]
3-OH-AAB	0.687 [9]	2.97	-1.02	^c
3-OCH ₂ CH ₂ OH-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	1.052 [10]	1.60	-0.89(-0.54)	132–134 [29]
3-OCH ₂ CH ₂ OH-AAB	1.348 [10]	2.51	-0.92(-0.49)	65–67 [29]
2'-CH ₂ OH-3-Me-AAB	2.012 [34]	2.40	-0.72	^c
4'-OMe-AAB	2.300 [30]	2.95	-1.09(-1.57)	155–159 [31]
2',3-diMe-AAB (OAT)	2.676 [34]	4.05	-2.00(-2.14)	101–102 [37]
3-OBu-AAB	4.983 [10]	5.08	-3.15(-2.92)	63–65 [29]
3-OEt-AAB	13.802 [10]	4.02	-2.07(-2.20)	107–109 [29]
3-OPr-AAB	18.919 [10]	4.55	-2.60(-2.61)	97–98 [29]
3-OMe-AAB	77.065 [10]	3.48	-1.54(-1.68)	110–111 [29]
<i>B. MAB (R=CH₃, R'=H)</i>				
3'-Me-4'-OH-MAB	0.071 [33]	3.67	-1.77	^c
3'-COOH-MAB	0.124 [33]	3.52	1.00(0.16)	207–208 [36]
4'-OH-MAB	0.140 [9]	3.21	-1.30	^c
MAB	0.183 [38]	3.79	-1.68(-1.74)	87.5–88 [38]
4'-Me-MAB	0.283 [38]	4.25	-2.16(-2.35)	105–105.5 [38]
3'-Me-MAB	0.445 [33]	4.25	-2.16(-2.39)	109–109.5 [32]
3'-CH ₂ OH-MAB	0.503 [33]	2.60	-0.89(-0.10)	119–121 [33]
<i>C. DAB (R=R'=CH₃)</i>				
3'-Me-4'-OH-DAB	0.110 [39]	4.31	-2.41	^c
DAB	0.140 [39]	4.43	-2.31(-2.62)	117–118 [32]
3'-COOH-DAB	0.201 [33]	4.17	0.37(-0.52)	212–213 [36]
2-Me-DAB	0.220 [39]	4.89	-2.80(-2.56)	64–66 [40]
3'-Me-DAB	0.356 [33]	4.89	-2.80(-3.52)	169–170 [36]
3'-CHO-DAB	0.383 [33]	3.91	-2.07(-2.06)	97–99 [36]
3'-CH ₂ OAc-DAB	0.518 [33]	4.14	-2.55(-2.13)	64–66 [36]
3'-CH ₂ OH-DAB	0.601 [33]	3.25	-1.52(-1.64)	122–123 [36]

(continued on next page)

Table 1 (continued)

Compounds	Mutagenic activity in TA98 + S9 (rev/nmol) [Ref.]	Log P^a	Log S^b	Melting point (°C) [Ref]
<i>D. Metabolites (R=OH, Ac; R'=H, CH₃)</i>				
3'-Me-AAB-N-Ac	0.087 [33]	3.73	-1.92	^c
3'-Me-4'-OH-AAB-N-Ac	0.089 [33]	3.15	-1.54(-2.13)	188 [36]
N-OH-2-OMe-AAB	0.110 [41]	4.08	-2.14(-2.59)	145–146 ^d [30]
3'-Me-MAB-N-Ac	0.524 [33]	3.03	-1.43(-1.70)	149–150 [36]
N-OH-MAB	0.650 [42]	2.74	-0.92(-1.51)	171–174 [38]
N-OH-3'-Me-MAB	1.000 [38]	3.20	-1.40	^c
N-OH-AAB	1.030 [41]	2.98	-1.03(-1.89)	195–197 [30]
N-OH-4'-Me-MAB	1.132 [38]	3.20	-1.40	^c
N-OH-3-OMe-AAB	192.000 [41]	3.08	-1.30(-1.38)	113–114 ^d [30]
<i>E. Potassium salts^e</i>				
3-OSO ₃ K-MAB	0.034 [9]	-1.05	2.38	^c
3-OSO ₃ K-AAB	0.037 [9]	-1.00	2.54	^c
4'-OSO ₃ K-AAB	0.097 [9]	-1.29	2.78	^c
4'-OSO ₃ K-MAB	0.422 [9]	-1.33	2.61	^c

^a These values of Log P are calculated at pH = 7 [24].

^b The values of S (g/l) were calculated at pH = 7. The values in parentheses were recalculated using the melting point data; this is expected to give better estimates of the aqueous solubility [24].

^c Melting points for these compounds could not be found.

^d Melting points of hydrochloride salts.

^e These compounds were optimized as ions.

descriptors that were considered in developing ANNs (vide infra).

In Table 4A we list BMLR equations for the relative mutagenic activity of the 74 compounds listed in Tables 1–3 using 4–8 descriptors; Fisher criterion F -values, F , variances, s^2 , and squares of the regression correlation coefficients, R^2 , are also given in this table. The descriptors in the correlation equations in Table 4A are identified in Table 5 [11,12,21,24,45–54]. The observed values of Log TA98 and the values predicted from the various BMLR models are given in Table 6. A correlation plot for the 8-descriptor model is shown in Fig. 1. Several of the descriptors that appear in one or more of the BMLR models are the same as those that appeared in our previous study [3]. For example, the electrostatic polarity parameter $(Q_{\text{MAX}} - Q_{\text{MIN}})/(\text{distance})^2$ [3,11] appears in both the 7- and 8-descriptor equations in Table 4A, and increasing the value of this parameter decreases the relative mutagenicity of the compound, similar to what we observed in our earlier study [3]. As would be expected, however,

the enhanced structural diversity of the compounds in the current data set required several new descriptors. For example, the relative negative charge, RNCG, appears in the 6-, 7-, and 8-descriptor BMLR equations in Table 4A. This electrostatic descriptor is defined as (QMNEG/QTMINUS), where QMNEG is the minimum negative atomic charge and QTMINUS is the sum of all the negative atomic charges [55]. Increasing the value of this parameter tends to increase the relative mutagenicity of the compound.

As can be seen from Table 4A, an 8-descriptor BMLR model accounts for about 73% of the variation in mutagenic activity of the compounds in Tables 1–3. For comparison, a 4-descriptor model for the monoazo derivatives in our previous study [3] was able to account for 79% of the variation. (It was necessary to use a total of 10 descriptors for the 74 compounds in this study to get a comparable R^2 value.) Although it is often difficult to identify the mechanistic significance of each molecular descriptor that appears in a QSAR/QPAR model, an analysis of the BMLR

equations in Table 4A reveals some interesting features. Each of these equations involves at least one descriptor associated with the nitrogen atom(s) in the molecule. The main descriptor in this regard is either the *minimum* or the *maximum* value of the electrophilic reactivity index for an N atom, ERI_N . ERI_N is defined as

$$ERI_N = \sum_{j \in N} \frac{C_{jLUMO}^2}{\epsilon_{LUMO} + 10}, \quad (1)$$

where C_{jLUMO} denotes the j th atomic orbital coefficient of the lowest unoccupied molecular orbital (LUMO) and ϵ_{LUMO} is the orbital energy of the LUMO [11]. In our previous study, in which all the compounds had both amino and azo nitrogen atoms, the *average* value of ERI_N was an important descriptor [3]. The value of ERI_N is a measure of the acidity of a nitrogen atom and, in general, larger values of this index occur at azo nitrogen atoms and smaller values occur at amino nitrogen atoms. The reductive cleavage compounds in this study, however, do not have an azo linkage, and it seems reasonable that the average

value of ERI_N has been replaced by the maximum or minimum value of this index. Since metabolic reductive-cleavage of the azo linkage and/or hydroxylation at the amino nitrogen atom (and subsequent conversion to an aryl nitrenium ion) are believed to be important steps in the onset of mutagenesis [2,8,56–58], the presence of descriptors based on ERI_N in these correlation equations seem appropriate. In the 4-6-descriptor equations, the mutagenicity increases as the minimum value of ERI_N increases, whereas in the 7- and 8-descriptor equations, the mutagenicity increases as the maximum value of ERI_N increases.

It is also quite striking that each of the BMLR equations in Table 4A includes the maximum total interaction of a C–H bond [11]; the mutagenicity decreases as the value of this index increases. Since detoxification mechanisms, such as C-hydroxylation, are believed to play a role in the relative mutagenicity of amino compounds, this descriptor may encode such features into the models.

The predictability of the BMLR models in Table 4 were assessed using a leave-one-out

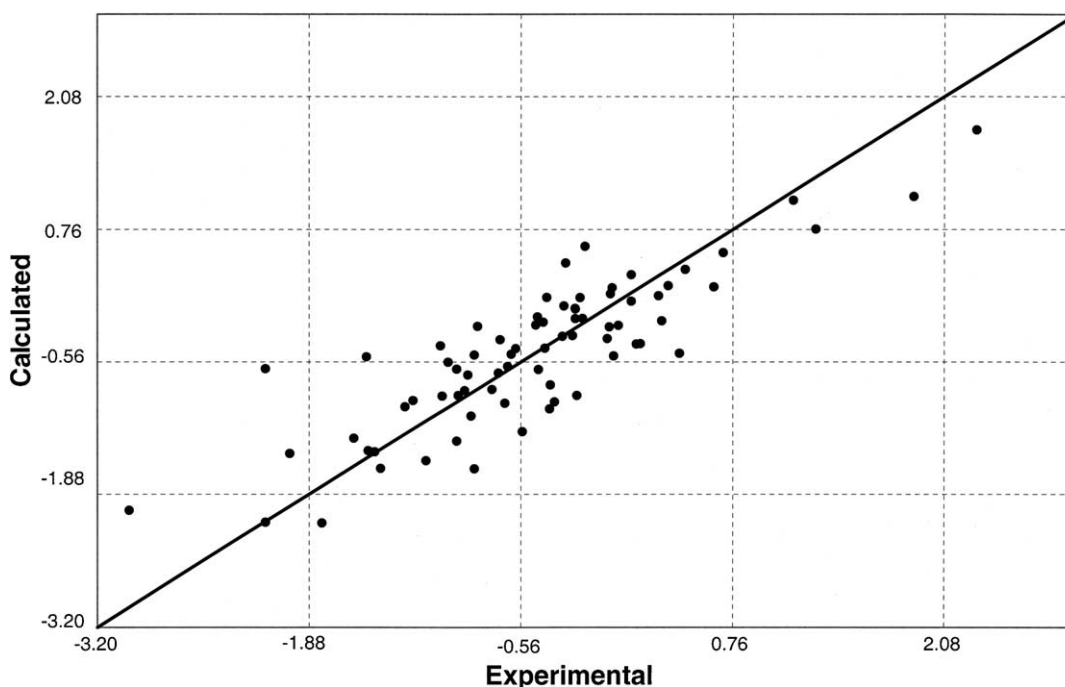
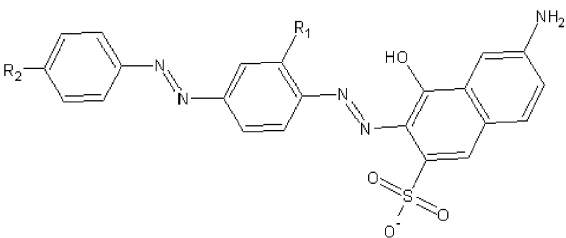


Fig. 1. Correlation plot for the 8-descriptor BMLR equation in Table 4.

Table 2

Observed mutagenicity (rev/nmol) in the TA98 *Salmonella typhimurium* bacterial strain with S9 activation, calculated values [24] of Log*P* and Log*S* for various disazo dyes



Compounds ^a	Mutagenic activity in TA98 + S9 (rev/nmol) [Ref.]	Log <i>P</i> ^b	Log <i>S</i> ^c
R ₁ = OCH ₂ CH ₂ OH, R ₂ = NEt ₂	0.148 [10]	0.12	−0.62
R ₁ = OBu, R ₂ = NEt ₂	0.228 [10]	2.69	−2.85
R ₁ = OPr, R ₂ = NEt ₂	0.256 [10]	2.16	−2.30
R ₁ = OMe, R ₂ = NEt ₂	0.342 [10]	1.09	−1.19
R ₁ = OCH ₂ CH ₂ OH, R ₂ = N(CH ₂ CH ₂ OH) ₂	0.350 [10]	−2.55	1.38
R ₁ = OPr, R ₂ = H	0.400 [10]	0.39	−0.27
R ₁ = OEt, R ₂ = NEt ₂	0.580 [10]	1.63	−1.74
R ₁ = OMe, R ₂ = N(CH ₂ CH ₂ OH) ₂	0.601 [10]	−1.58	0.80
R ₁ = OBu, R ₂ = H	0.668 [10]	0.92	−0.82
R ₁ = OEt, R ₂ = N(CH ₂ CH ₂ OH) ₂	0.983 [10]	−1.05	0.25
R ₁ = OCH ₂ CH ₂ OH, R ₂ = H	1.348 [10]	−1.65	1.42
R ₁ = OBu, R ₂ = N(CH ₂ CH ₂ OH) ₂	1.452 [10]	0.01	−0.85
R ₁ = OPr, R ₂ = N(CH ₂ CH ₂ OH) ₂	1.533 [10]	−0.52	−0.30
R ₁ = OMe, R ₂ = H	2.920 [10]	−0.67	0.83
R ₁ = OEt, R ₂ = H	4.383 [10]	−0.14	0.28

^a These compounds were optimized as ions.

^b These values of Log*P* are calculated at pH = 7 [24].

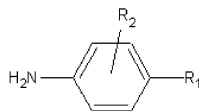
^c The values of *S* (g/l) were calculated at pH = 7. Melting points for these compounds could not be found in the literature.

(LOO) cross-validation procedure [11,59]. For each of the 74 compounds, multilinear regression was recalculated for the data set without this compound, but using the same descriptors. The resulting regression equation, based on 73 compounds, was then used to predict the value of the LogTA98 for the compound that was removed. The set of predicted values of LogTA98 was linearly correlated with the array of experimental values; the cross-validated correlation coefficient, (*R*_{CV})², was calculated and the values are listed in Table 4A for each of our BMLR models. As can be seen from this Table, the values of (*R*_{CV})² are typically about 0.1 lower than the value of *R*², suggesting reasonable generalizability of these models.

It is interesting to note that the hydrophobicity descriptor Log*P* does not appear in any of the BMLR correlation equations in Table 4A or in the previous BMLR equations we developed that were limited to AAB, MAB, and DAB derivatives [3]. Furthermore, when we used the heuristic method implemented in CODESSA [11] to force Log*P* to be one of the descriptors in various BMLR models, it did not give any improvement over other derived models, see Table 4B. This is somewhat surprising since the QSAR studies of Debnath et al. [22] suggest that hydrophobicity plays an important role in regulating mutagenic activity in TA98 + S9 for a variety of aromatic and heteroaromatic amines. However, in a more recent comprehensive QSAR treatment of the Debnath et al. [22] data

Table 3

Observed mutagenicity (rev/nmol) in the TA98 *Salmonella typhimurium* bacterial strain with S9 activation, calculated values [24] of Log*P*, Log*S*, and melting points (°C) for various reductive cleavage products of the aminoazobenzene dyes and disazo dyes in Tables 1 and 2



Compounds	Mutagenic activity in TA98 + S9 (rev/nmol) [Ref.]	Log <i>P</i> ^a	Log <i>S</i> ^b	Melting point (°C) [Ref.]
2-OBu-1,4-phenylenediamine	0.001 [10]	0.89	0.99	185 ^c [10]
NEt ₂ -4-phenylenediamine	0.007 [10]	1.81	1.47	184–186 ^d [43]
4- <i>N</i> -Acetylamino-aniline	0.016 [9]	0.08	1.77	^e
4-OH-aniline	0.025 [9]	−0.29	2.29	188–190 [43]
2-Me-1,4-phenylenediamine	0.030 [9]	−0.39	2.36	64 [43]
Aniline	0.031 [9]	0.94	1.33	−6 [43]
1,4-Phenylenediamine	0.129 [9]	−0.85	2.81	143–145 [43]
NMe ₂ -4-phenylenediamine	0.136 [2]	0.75	1.45	^e
4-Methylamino-aniline	0.414 [9]	−0.09	2.36	^e
2-OPr-1,4-phenylenediamine	0.619 [10]	0.36	1.51	202 ^c [10]
2-OEt-1,4-phenylenediamine	0.963 [10]	−0.18	2.00	210 ^c [10]
2-OMe-1,4-phenylenediamine	2.072 [10]	−0.71	1.53	^e

^a These values of Log*P* are calculated at pH = 7 [24].

^b The values of *S* (g/L) were calculated at pH = 7.

^c Melting points of hydrochloride salts.

^d Melting points of sulfate salts.

^e Melting points for these compounds could not be found.

Table 4

Quantitative structure–activity/structure–property relationships for the mutagenicity of the compounds in Tables 1–3^a

QSAR/QPAR	<i>N</i>	<i>R</i> ²	<i>s</i> ²	<i>F</i>	(<i>R</i> _{cv}) ²
<i>A. BMLR equations</i>					
Log TA98 = 591.96(±91.34)−18.76(±3.67)Q ₁ −25.7(±4.40)Q ₂ + 929.93(±162.69)Q ₃ −53.80(±12.64)Q ₄	4	0.47	0.42	15.30	0.38
Log TA98 = 439.33(±58.98)−11.26(±3.83)Q ₁ −29.65(±4.39)Q ₂ + 1610.10(±241.89)Q ₃ −7.32(±1.48)Q ₄ −1.36(±0.43)T ₁	5	0.53	0.38	15.51	0.43
Log TA98 = 402.50(±52.49) + 2.94(±4.60)Q ₁ −29.84(±4.06)Q ₂ + 2179.80(±269.15)Q ₃ −9.32(±1.34)Q ₄ + 17.85(±3.60)E ₁ + 38.76(±8.33)E ₂	6	0.62	0.31	18.29	0.50
Log TA98 = 304.72(±38.80)−15.91(±2.78)E ₃ + 274.62(±27.48) Q ₅ −24.86(±3.10)Q ₂ + 23.85(±3.33)E ₁ −4.17(±0.68)E ₄ −0.02(±0.00)E ₅ −0.01(±0.00)H ₁	7	0.67	0.27	19.55	0.59
Log TA98 = 271.99(±32.30)−9.63(±2.88)E ₃ + 251.05(±22.53)Q ₅ −22.30 (±2.59)Q ₂ + 32.66(±3.88)E ₁ −3.55(±0.65)E ₄ −12.74(±1.66)E ₆ −12.58 (±3.22)G ₁ −64.30(±14.73)E ₇	8	0.73	0.23	21.53	0.64
<i>B. Heuristic equations</i>					
log TA98 = 344.37(±60.25) + 3.05(±1.88)Q ₆ −26.72(±4.63)Q ₂ −0.10 (±0.10)O ₁ + 1219.10(±285.97)Q ₃ −3.04(±1.10)Q ₇	5	0.44	0.45	10.65	0.33
Log TA98 = 302.15(±55.51) + 64.38(±19.4)Q ₅ −23.82(±4.39)Q ₂ −0.13 (±0.06)O ₁ + 826.56(±200.52)Q ₃ −4.47(±1.64)G ₂	5	0.44	0.45	10.61	0.33

^a The molecular descriptors in these BMLR equations are identified in Table 5.

Table 5
Molecular descriptors for the QSAR/QPAR models in Table 4

Classification/label	Molecular descriptors	Reference
<i>Geometrical</i>		
G ₁	Principle moment of inertia A	11
G ₂	XY Shadow/XY Rectangle	11, 45
<i>Electrostatic</i> ^a		
E ₁	RNCG Relative negative charge (QMNEG/QTMINUS)	11
E ₂	Minimum partial charge for a N atom	11
E ₃	Minimum partial charge (Q_{MIN})	11
E ₄	Polarity parameter/(distance) ² = ($Q_{\text{MAX}} - Q_{\text{MIN}}$)/(distance) ²	11, 46, 47
E ₅	HACA H-acceptors charged surface area	11
E ₆	FHBCA fractional HBSA (HBSA/TMSA) ^d	11, 48, 49
E ₇	FNSA-3 fractional PNSA (PNSA-3/TMSA) ^d	11, 48, 49
<i>Quantum-chemical</i>		
Q ₁	Average valency of a N atom	11
Q ₂	Maximum total interaction for a C–H bond	11
Q ₃	Minimum electrophilic reactivity index for a N atom	11, 50
Q ₄	Average valency of a C atom	11
Q ₅	Maximum electrophilic reactivity index for a N atom	11, 50
Q ₆	Average bond order of a N atom	11
Q ₇	Minimum valency of a C atom	11
<i>Thermodynamic</i>		
H ₁	Final heat of formation	11, 12
<i>Topological</i> ^b		
T ₁	Average complementary information content (order 0) ^c	11
<i>Others</i>		
O ₁	LogP	21, 24

^a These descriptors reflect characteristics of the charge distribution of the molecule. The empirical partial charges in a molecule are calculated using the approach of Zefirov et al. [46,47] which is based on the Sanderson electronegativity scale [51,52].

^b Topological indices describe the atomic connectivity in the molecule.

^c The average information content is defined on the basis of Shannon information theory [53,54] and is calculated as

$$IC = -\sum_i (n_i/n) \log_2(n_i/n)$$

where n_i is the number of atoms in the i th class and n is the total number of atoms in the molecule. The division of atoms into different classes depends upon the coordination sphere taken into account.

^d TMSA is the total molecular surface area; PNSA-3 is the atomic charge weighted partial negative surface area; HBSA is the hydrogen bonding surface area.

set, Maran et al. [60] found that LogP did not appear as a descriptor in their BMLR equations.

5. Artificial neural networks

Based on our regression results described above, and the experience of other authors in mutageni-

city studies [22, 60], we selected a pool of about 50 descriptors to construct our neural networks. Although an exhaustive search was not practical, a large variety of networks with different numbers and types of descriptors were generated and evaluated. In Table 7 we describe 5 ANNs for the relative mutagenic activity in TA98 + S9 of the 74 compounds listed in Tables 1–3; networks I–V use

Table 6

Observed values of LogTA98 and predicted values of LogTA98 from the BMLR models in Table 4 for the 74 compounds in Tables 1–3

Compounds	Observed values of LogTA98	Predicted values of LogTA98				
		BMLR				
		4-Descriptor	5-Descriptor	6-Descriptor	7-Descriptor	8-Descriptor
<i>A. AAB</i>						
4'-NEt ₂ -OMe-AAB	-2.15	-0.58	-1.08	-1.11	-0.29	-0.63
2-OMe-AAB	-2.00	-1.39	-1.42	-1.17	-1.55	-1.47
4'-OH-AAB	-1.28	-0.65	-0.73	-0.84	-0.79	-1.01
3'-Me-4'-OH-AAB	-1.23	-0.52	-0.46	-0.69	-0.74	-0.95
4'-OH-2',3-diMe-AAB (4'-OH-OAT)	-0.95	-0.57	-0.55	-0.85	-0.95	-0.90
AAB	-0.69	-0.50	-0.44	0.03	-0.44	-0.34
3'-Me-AAB	-0.62	-0.56	-0.40	-0.13	-0.71	-0.48
3-OMe-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	-0.41	-0.16	-0.84	-0.89	-0.39	-0.42
3'-CH ₂ OH-AAB	-0.22	-0.48	-0.25	-0.32	-0.37	-0.03
3-OH-AAB	-0.16	0.40	0.87	1.05	0.58	0.59
3-OCH ₂ CH ₂ OH-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	+0.02	-0.15	-1.06	-1.13	-0.42	-0.50
3-OCH ₂ CH ₂ OH-AAB	+0.13	0.44	0.72	0.71	-0.22	0.05
2'-CH ₂ OH-3-Me-AAB	+0.30	-0.26	0.14	0.18	-0.29	0.10
4'-OMe-AAB	+0.36	-0.14	-0.10	-0.15	0.71	0.20
2',3-diMe-AAB (OAT)	+0.43	-0.38	-0.11	0.22	-0.83	-0.47
3-OBu-AAB	+0.70	0.12	0.26	0.32	0.76	0.53
3-OEt-AAB	+1.14	0.24	0.56	0.68	1.19	1.05
3-OPr-AAB	+1.28	0.22	0.50	0.60	0.95	0.76
3-OMe-AAB	+1.89	0.09	0.50	0.88	1.05	1.09
<i>B. MAB</i>						
3'-Me-4'-OH-MAB	-1.15	-0.77	-0.82	-1.24	-1.26	-1.55
3'-COOH-MAB	-0.91	-0.44	-0.36	-0.67	-1.13	-0.85
4'-OH-MAB	-0.85	-0.90	-0.99	-1.43	-1.34	-1.63
MAB	-0.74	-0.53	-0.42	-0.35	-0.87	-0.83
4'-Me-MAB	-0.55	-0.67	-0.64	-0.69	-1.14	-1.26
3'-Me-MAB	-0.35	-0.59	-0.48	-0.47	-1.08	-0.96
3'-CH ₂ OH-MAB	-0.30	-0.53	-0.51	-0.66	-0.46	-0.30
<i>C. DAB</i>						
3'-Me-4'-OH-DAB	-0.95	-0.27	-0.31	-0.37	-0.40	-0.49
DAB	-0.85	-0.32	-0.42	-0.93	-1.18	-1.35
3'-COOH-DAB	-0.70	-0.31	-0.34	-0.46	-0.96	-0.67
2-Me-DAB	-0.66	-0.66	-0.81	-0.93	-0.97	-0.97
3'-Me-DAB	-0.45	-0.41	-0.45	-0.54	-0.60	-0.63
3'-CHO-DAB	-0.42	-0.22	-0.24	-0.41	-0.63	-0.16
3'-CH ₂ OAc-DAB	-0.29	0.13	0.09	0.08	0.07	0.00
3'-CH ₂ OH-DAB	-0.22	-0.32	-0.43	-0.33	-0.33	-0.13
<i>D. Metabolites</i>						
3'-Me-AAB-N-Ac	-1.06	-1.53	-1.55	-0.97	-0.36	-0.40
3'-Me-4'-OH-AAB-N-Ac	-1.05	-1.35	-1.45	-1.20	-0.72	-0.90
N-OH-2-OMe-AAB	-0.96	-0.84	-1.37	-0.96	-0.69	-0.63
3'-Me-MAB-N-Ac	-0.28	-0.94	-1.22	-0.69	0.21	0.43
N-OH-MAB	-0.19	0.38	0.06	0.40	0.25	0.08
N-OH-3'-Me-MAB	+0.00	0.45	0.13	0.44	0.26	0.12

(continued on next page)

Table 6 (continued)

Compounds	Observed values of LogTA98	Predicted values of LogTA98				
		BMLR				
		4-Descriptor	5-Descriptor	6-Descriptor	7-Descriptor	8-Descriptor
N-OH-AAB	+0.01	0.28	−0.04	0.50	0.22	0.18
N-OH-4'-Me-MAB	+0.05	0.43	0.06	0.35	0.12	−0.19
N-OH-3-OMe-AAB	+2.28	0.75	0.58	1.11	1.82	1.75
<i>E. Potassium salts</i>						
3-OSO ₃ K-MAB	−1.47	−1.05	−1.14	−1.35	−1.04	−1.46
3-OSO ₃ K-AAB	−1.43	−0.93	−1.05	−1.43	−1.20	−1.62
4'-OSO ₃ K-AAB	−1.01	−0.84	−0.75	−0.99	−0.54	−0.56
4'-OSO ₃ K-MAB	−0.38	−1.12	−1.12	−1.31	−0.57	−0.79
<i>F. Disazo dyes</i>						
R ₁ = OCH ₂ CH ₂ OH, R ₂ = NEt ₂	−0.83	−0.20	−0.22	−0.23	−0.30	−0.20
R ₁ = OBu, R ₂ = NEt ₂	−0.64	−0.37	−0.44	−0.30	−0.52	−0.60
R ₁ = OPr, R ₂ = NEt ₂	−0.59	−0.30	−0.31	−0.22	−0.48	−0.43
R ₁ = OMe, R ₂ = NEt ₂	−0.47	−0.22	−0.14	−0.12	−0.47	−0.19
R ₁ = OCH ₂ CH ₂ OH, R ₂ = N(CH ₂ CH ₂ OH) ₂	−0.46	−0.09	−0.27	−0.31	−0.15	−0.11
R ₁ = OPr, R ₂ = H	−0.40	−0.17	0.12	0.05	0.01	0.08
R ₁ = OEt, R ₂ = NEt ₂	−0.24	−0.24	−0.21	−0.17	−0.46	−0.30
R ₁ = OMe, R ₂ = N(CH ₂ CH ₂ OH) ₂	−0.22	−0.07	−0.15	−0.20	−0.22	−0.02
R ₁ = OBu, R ₂ = H	−0.18	−0.23	0.03	−0.01	−0.03	−0.13
R ₁ = OEt, R ₂ = N(CH ₂ CH ₂ OH) ₂	−0.01	−0.13	−0.24	−0.26	−0.31	−0.21
R ₁ = OCH ₂ CH ₂ OH, R ₂ = H	+0.13	−0.05	0.19	0.04	0.19	0.31
R ₁ = OBu, R ₂ = N(CH ₂ CH ₂ OH) ₂	+0.16	−0.26	−0.44	−0.35	−0.31	−0.38
R ₁ = OPr, R ₂ = N(CH ₂ CH ₂ OH) ₂	+0.19	−0.26	−0.44	−0.35	−0.32	−0.38
R ₁ = OMe, R ₂ = H	+0.47	−0.03	0.28	0.19	0.10	0.36
R ₁ = OEt, R ₂ = H	+0.64	−0.10	0.21	0.11	0.06	0.19
<i>G. Reductive-cleavage products/metabolites</i>						
2-OBu-1-4-phenylenediamine	−3.00	−0.87	−1.16	−1.95	−1.76	−2.04
N-Et ₂ -phenylenediamine	−2.15	−1.09	−1.42	−1.50	−2.33	−2.16
4-N-Acetylamino-aniline	−1.80	−2.12	−1.51	−1.16	−1.71	−2.17
4-OH-aniline	−1.60	−1.42	−1.07	−1.14	−0.96	−1.32
2-Me-1-4-phenylenediamine	−1.52	−0.94	−0.56	−0.82	−0.99	−0.51
Aniline	−1.51	−2.88	−2.81	−1.98	−1.60	−1.45
1-4-Phenylenediamine	−0.89	−0.91	−0.78	−0.94	−0.46	−0.69
N-Me ₂ -4-phenylenediamine	−0.87	−1.22	−1.42	−1.23	−1.17	−1.10
4-Methylamino-aniline	−0.38	−0.96	−0.75	−0.70	−0.80	−1.03
2-OPr-1-4-phenylenediamine	−0.21	−0.55	−0.55	−1.25	−1.10	−0.89
2-OEt-1-4-phenylenediamine	−0.02	−0.43	−0.29	−0.93	−0.72	−0.33
2-OMe-1-4-phenylenediamine	0.32	−0.36	−0.14	−0.56	−0.29	−0.15

from 4 to 8 molecular descriptors, respectively. For each ANN, we list the descriptors involved, their classification and relative importance to the network, and the value of the correlation coefficient, R^2 . The observed values of LogTA98 and the values predicted from the various ANNs are given

in Table 8, and a correlation plot for the 8-descriptor case is shown in Fig. 2.

It is well known that ANNs with a single hidden layer, containing a sufficient number of neurons, can interpolate any multidimensional nonlinear function to a given accuracy, and can implement

exactly an arbitrary finite training set [61]. This remarkable potential of ANNs to pattern non-linear phenomena, however, can easily result in relatively poor predictive power of a trained ANN toward new data. To evaluate the generalization ability of our various networks, an *n*-fold cross-validation procedure was employed [63], in which the 74 compounds were split into 10 subsets containing either 7 or 8 compounds. The network was then retrained 10 times, using the same descriptors, but each time leaving out one of the subsets from the training; values of LogTA98 were predicted from the resulting ANN for the compounds left out and the correlation coefficient, R^2 , was calculated. The average of the ten values of R^2 was used as the cross-validation correlation coefficient, $(R_{cv})^2$. ANNs I–V in Table 8 were chosen from all the networks we generated based on the values of both R^2 and $(R_{cv})^2$.

The molecular descriptors used in these ANNs are often different from those in the various BMLR equations in Table 4A. For example, the total dipole moment and various polarizabilities derived from quantum chemistry play important roles in the neural networks in Table 7, but are not

involved in the BMLR equations. Furthermore, each of the networks in Table 7 involves at least one shape index [45], e.g. the XY shadow/XY rectangle, whereas such geometrical indices played relatively minor roles in only two of the regression equations in Table 4. Interestingly, Log*P* is an integral component of several of the ANNs we developed, although its relative importance is quite low for the 6- and 8-descriptor cases, and it was not included in the 7-descriptor ANN.

The 8-descriptor ANN in Table 7 accounts for some 95% of the reported variation in the relative mutagenicity of the 74 compounds in Tables 1–3; the value of $(R_{cv})^2$ is also extremely good, 0.94, suggesting this ANN also has excellent predictive power. As can be seen from Fig. 2, this network fails to predict accurately values of LogTA98 in only 4 instances: 3'-Me-4'-OH-AAB, 4-OH-Aniline, 4'-Me-MAB, and 3'-Me-DAB. The most serious discrepancy involves the dye 3'-Me-4'-OH-AAB, which is predicted to show substantial mutagenicity, LogTA98 = +0.36, whereas the observed value is much lower, −1.23. It is difficult to understand the reason

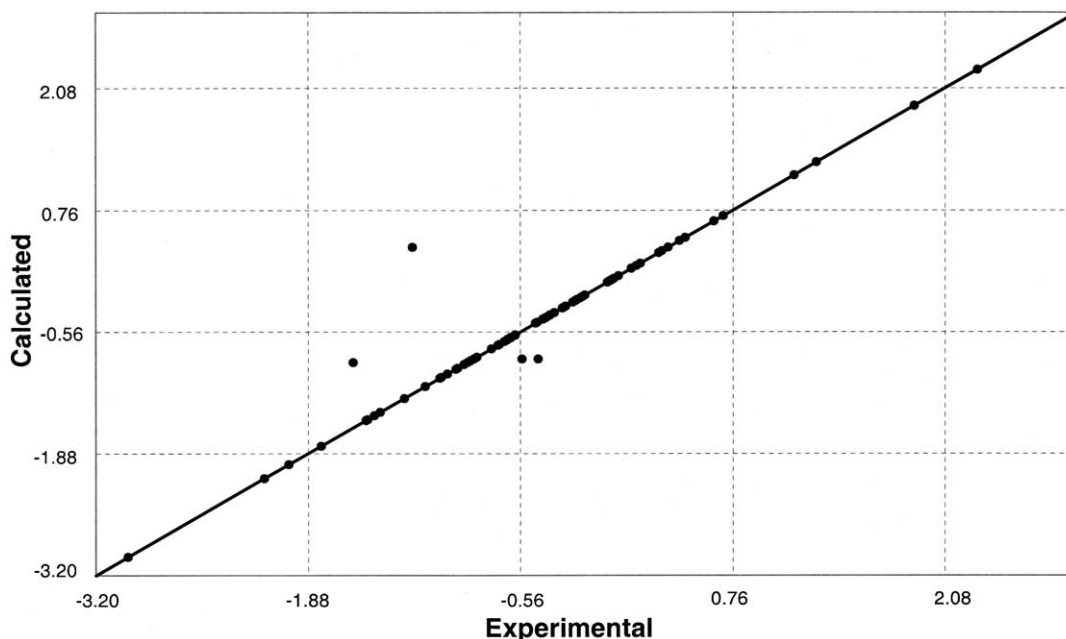


Fig. 2. Correlation plot for the 8-descriptor ANN (V) in Table 7.

Table 7

Molecular descriptors and their relative importance (%) for several artificial neural networks [25]

Model	Molecular descriptors	Classification ^a	Relative importance (%)	<i>N</i>	<i>R</i> ²	(<i>R</i> _{Cv}) ²
I	Log <i>P</i>	O	22.6	4	0.66	0.61
	Total dipole of the molecule	Q	17.4			
	XY Shadow/XY rectangle ^c	G	20.3			
	Balaban index ^b	T	39.7			
II	Log <i>P</i>	O	42.4	5	0.77	0.68
	Number of benzene rings	C	1.8			
	XY Shadow/XY rectangle ^c	G	2.4			
	1x Gamma polarizability	Q	49.2			
	1/2x Beta polarizability	Q	4.2			
III	Log <i>P</i>	O	2.1	6	0.82	0.78
	Number of benzene rings	C	0.2			
	XY Shadow/XY rectangle ^c	G	4.6			
	Minimum valency of a C atom	Q	32.7			
	Average valency of a N atom	Q	36.6			
	1x Gamma polarizability	Q	26.9			
IV	Total dipole of the molecule	Q	0.1	7	0.91	0.89
	Average valency of a C atom	Q	47.6			
	Min valency of a C atom	Q	45.2			
	1x Gamma polarizability	Q	4.1			
	1/2x Beta polarizability	Q	1.9			
	Internal entropy/number of atoms	H	0.4			
	YZ Shadow/YZ rectangle ^c	G	0.7			
V	Log <i>P</i>	O	4.2	8	0.95	0.94
	Number of benzene rings	C	16.8			
	XY Shadow/XY rectangle ^c	G	5.5			
	1x Gamma polarizability	Q	6.9			
	Balaban Index ^b	T	53.5			
	YZ Shadow/YZ rectangle ^c	G	4.5			
	1/2x Beta polarizability	Q	2.1			
	Total dipole of the molecule	Q	6.6			

^a C: Constitutional; T: Topological; E: Electrostatic; G: Geometrical; Q: Quantum-Chemical; H: Thermodynamic; O: Other.^b The Balaban index [62] is defined by

$$J = (q/\mu + 1) \sum_{i,j}^q (s_i \bullet s_j)^{-1/2}$$

where *q* is the number of edges in the molecular graph, *n* is the number of vertices in the graph, $\mu = q - n + 1$ is the cyclometric number, and *s_i* is the distance sums obtained by summing the *i*th row and *i*th column of the distance matrix between atoms in the molecule [11].

^c This is one of the six indices defined by Rohrbaugh and Jurs [45] that encode size and shape information about a molecule in three mutually perpendicular perspectives defined by the axes of inertia.

for this. The experimental value of LogTA98 seems quite reasonable in relation to the corresponding experimental values for AAB, 3'-Me-AAB, and 4'-OH-AAB, see Table 1. Furthermore,

the 7-descriptor ANN does predict a value of −1.23 for 3'-Me-4'-OH-AAB and the 8-descriptor BMLR model gives a reasonable value of −0.95.

Table 8

Observed values of LogTA98 and predicted values of LogTA98 from the ANNs for the 74 compounds in Tables 1–3

Compounds	Observed values of LogTA98	Predicted values of LogTA98 ANN				
		I	II	III	IV	V
		4-Descriptor	5-Descriptor	6-Descriptor	7-Descriptor	8-Descriptor
<i>A. AAB</i>						
4'-NEt ₂ -3-OMe-AAB	-2.15	-2.15	-2.15	-2.15	-2.15	-2.15
2-OMe-AAB	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00
4'-OH-AAB	-1.28	-0.85	-0.85	-1.28	-1.28	-1.28
3'-Me-4'-OH-AAB	-1.23	-0.22	-0.95	-0.95	-1.23	+0.36
4'-OH-2',3-diMe-AAB (4'-OH-OAT)	-0.95	-0.95	-0.95	-0.95	-0.95	-0.95
AAB	-0.69	-0.85	+0.30	-0.69	-0.69	-0.69
3'-Me-AAB	-0.62	-0.95	-0.95	+0.43	-0.62	-0.62
3-OMe-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	-0.41	-0.41	-1.47	-0.41	-0.41	-0.41
3'-CH ₂ OH-AAB	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
3-OH-AAB	-0.16	+0.00	+0.00	-0.16	-0.16	-0.16
3-OCH ₂ CH ₂ OH-4'-N(CH ₂ CH ₂ OH) ₂ -AAB	+0.02	+0.02	+0.02	+0.02	+0.02	+0.02
3-OCH ₂ CH ₂ OH-AAB	+0.13	+0.13	+0.13	+0.13	+0.13	+0.13
2'-CH ₂ OH-3-Me-AAB	+0.30	-2.15	+0.30	+0.30	+0.30	+0.30
4'-OMe-AAB	+0.36	-0.22	-0.95	-0.95	+0.36	+0.36
2',3-diMe-AAB (OAT)	+0.43	-0.95	+0.43	+0.43	+0.30	+0.43
3-OBu-AAB	+0.70	+0.70	+0.70	+0.70	+0.70	+0.70
3-OEt-AAB	+1.14	+1.28	+1.28	+1.14	+1.14	+1.14
3-OPr-AAB	+1.28	+1.28	+1.28	+1.28	+1.28	+1.28
3-OMe-AAB	+1.89	+2.28	+1.89	+2.28	+1.89	+1.89
<i>B. MAB</i>						
3'-Me-4'-OH-MAB	-1.15	-0.22	-0.95	-0.95	-1.15	-1.15
3'-COOH-MAB	-0.91	-0.91	-0.91	-0.91	-0.91	-0.91
4'-OH-MAB	-0.85	-0.22	-0.95	-0.95	-0.85	-0.85
MAB	-0.74	-0.95	-0.22	-0.22	-0.74	-0.74
4'-Me-MAB	-0.55	-0.95	-0.95	-0.95	-0.55	-0.85
3'-Me-MAB	-0.35	-0.95	-0.95	-0.35	-0.35	-0.35
3'-CH ₂ OH-MAB	-0.30	-0.22	-0.22	-0.22	-0.30	-0.30
<i>C. DAB</i>						
3'-Me-4'-OH-DAB	-0.95	-0.95	-0.95	-0.95	-0.95	-0.95
DAB	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
3'-COOH-DAB	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
2-Me-DAB	-0.66	-0.66	-0.66	-0.45	-0.66	-0.66
3'-Me-DAB	-0.45	-0.66	-0.85	-0.45	-0.45	-0.85
3'-CHO-DAB	-0.42	-0.42	-0.42	-0.95	-0.42	-0.42
3'-CH ₂ OAc-DAB	-0.29	+1.28	-0.29	-0.29	-0.29	-0.29
3'-CH ₂ OH-DAB	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
<i>D. Metabolites</i>						
3'-Me-AAB-N-Ac	-1.06	-0.22	-1.06	-1.06	-1.06	-1.06
3'-Me-4'-OH-AAB-N-Ac	-1.05	-0.22	-1.05	-1.05	-1.05	-1.05
N-OH-2-OMe-AAB	-0.96	+1.14	-0.96	-0.96	-0.96	-0.96
3'-Me-MAB-N-Ac	-0.28	-0.95	-0.28	-0.28	-0.28	-0.28
N-OH-MAB	-0.19	-0.19	-0.19	-0.19	-0.19	-0.19
N-OH-3'-Me-MAB	+0.00	+0.01	+0.00	+0.00	+0.00	+0.00
N-OH-AAB	+0.01	+0.01	+0.01	+0.01	+0.01	+0.01

(continued on next page)

Table 8 (continued)

Compounds	Observed values of LogTA98	Predicted values of LogTA98 ANN				
		I	II	III	IV	V
		4-Descriptor	5-Descriptor	6-Descriptor	7-Descriptor	8-Descriptor
N-OH-4'-Me-MAB	+0.05	+0.05	+0.05	+0.05	+0.05	+0.05
N-OH-3-OMe-AAB	+2.28	+2.28	+2.28	+2.28	+2.28	+2.28
<i>E. Potassium salts</i>						
3-OSO ₃ K-MAB	-1.47	-1.47	-1.43	-1.47	-1.47	-1.47
3-OSO ₃ K-AAB	-1.43	-1.43	-1.43	-1.43	-1.43	-1.43
4'-OSO ₃ K-AAB	-1.01	-1.01	-1.01	-1.01	-1.01	-1.01
4'-OSO ₃ K-MAB	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38
<i>F. Disazo dyes</i>						
R ₁ = OCH ₂ CH ₂ OH, R ₂ = NEt ₂	-0.83	-0.83	-0.83	-0.83	-0.59	-0.83
R ₁ = OBU, R ₂ = NEt ₂	-0.64	-0.64	-0.64	-0.64	-0.64	-0.64
R ₁ = OPr, R ₂ = NEt ₂	-0.59	-0.64	-0.64	-0.64	-0.59	-0.59
R ₁ = OMe, R ₂ = NEt ₂	-0.47	-0.47	-0.47	-0.47	-0.47	-0.47
R ₁ = OCH ₂ CH ₂ OH, R ₂ = N(CH ₂ CH ₂ OH) ₂	-0.46	-0.46	-0.46	-0.46	-0.46	-0.46
R ₁ = OPr, R ₂ = H	-0.40	-0.18	-0.18	-0.18	-0.40	-0.40
R ₁ = OEt, R ₂ = NEt ₂	-0.24	-0.24	-0.24	-0.24	-0.64	-0.24
R ₁ = OMe, R ₂ = N(CH ₂ CH ₂ OH) ₂	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
R ₁ = OBU, R ₂ = H	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18
R ₁ = OEt, R ₂ = N(CH ₂ CH ₂ OH) ₂	-0.01	-0.22	-0.01	-0.01	-0.46	-0.01
R ₁ = OCH ₂ CH ₂ OH, R ₂ = H	+0.13	+0.13	+0.13	+0.13	+0.13	+0.13
R ₁ = OBU, R ₂ = N(CH ₂ CH ₂ OH) ₂	+0.16	+0.16	+0.16	+0.16	-0.01	+0.16
R ₁ = OPr, R ₂ = N(CH ₂ CH ₂ OH) ₂	+0.19	+0.19	+0.19	+0.19	-0.01	+0.19
R ₁ = OMe, R ₂ = H	+0.47	+0.47	+0.47	+0.47	+0.47	+0.47
R ₁ = OEt, R ₂ = H	+0.64	-0.18	-0.18	-0.18	+0.64	+0.64
<i>G. Reductive-cleavage products/metabolites</i>						
2-OBu-1-4-phenylenediamine	-3.00	-3.00	-3.00	-3.00	-3.00	-3.00
N-Et ₂ -phenylenediamine	-2.15	-2.15	-2.15	-2.15	-2.15	-2.15
4-N-Acetylamino-aniline	-1.80	-1.80	-1.80	-1.80	-1.80	-1.80
4-OH-aniline	-1.60	+0.32	+0.32	-0.89	-1.60	-0.89
2-Me-1-4-phenylenediamine	-1.52	-1.52	-1.52	-1.52	-1.52	-1.52
Aniline	-1.51	-1.51	-1.51	-1.51	-1.51	-1.51
1-4-Phenylenediamine	-0.89	-0.89	-0.89	-0.89	-0.89	-0.89
N-Me ₂ -4-phenylenediamine	-0.87	-0.87	-0.87	-0.87	-0.87	-0.87
4-Methylamino-aniline	-0.38	-0.38	-1.52	-0.38	-0.38	-0.38
2-OPr-1-4-phenylenediamine	-0.21	-0.21	-0.21	-0.21	-3.00	-0.21
2-OEt-1-4-phenylenediamine	-0.02	-0.38	-1.52	-1.52	-0.02	-0.02
2-OMe-1-4-phenylenediamine	+0.32	+0.32	-0.89	+0.32	+0.32	+0.32

6. Concluding remarks

QSAR/QPAR approaches have been used extensively to study many types of biological activity, including chemical mutagenicity/carcinogenicity of a variety of different classes of compounds, e.g., nitroarenes [64]. The use of these

approaches in dealing with problems of azo dye genotoxicity, however, have been rather limited [3, 65–68]. In view of the widespread use of azo dyes in various consumer products [69] and their novel utilization in the synthesis of polymer catalysts and chemical sensors [70], developing practical, reliable QSAR/QPARs to screen new and existing

azo dyes, as well as, their metabolic and reductive-cleavage products, for potential toxicological problems is of vital importance. In this investigation we developed QSAR/QPARs for the observed mutagenicity in the *Salmonella typhimurium* TA98 bacterial tester strain (+S9) for a set of 74 compounds containing an amino subgroup—62 of the compounds also contain at least one azo linkage and 12 are their reductive-cleavage products. Many properties of the compounds in this data set are quite diverse, e.g., their average value of $\text{Log}P$ is 2.03 ± 2.03 ; for comparison, the average value of $\text{Log}P$ for the compounds in our earlier study was much higher, 3.51, and the standard deviation was much lower, 0.83, indicative of their more homogeneous nature [3].

We have shown that an 8-descriptor BMLR equation can account for about 73% of the variation in relative mutagenic activity of the 74 compounds in our data set. An examination of trends in many of the calculated molecular descriptors for the compounds in this study versus their observed mutagenic activity, however, strongly indicates rather complex nonlinear behavior, which is typical of biological phenomena. The special interest in ANNs for QSAR/QPAR development arises from their ability to cope with nonlinear relationships in the absence of an underlying model. Indeed, we have shown that an 8-descriptor ANN, developed with the novel FCID3 algorithm implemented by Cios and Sztandera [28], can account for 95% of the variation in the mutagenic activity of the compounds in this data set. Furthermore, the predictive power of this network, as assessed by cross-validation, is exceptionally good, $(R_{CV})^2 = 0.94$.

Much remains to be done, however, in developing more reliable genotoxicity QSAR/QPARs for azo derivatives and their various metabolic products. The paucity of quantitative mutagenicity data for more structurally diverse azo dyes severely limits the range of compounds that can be screened reliably using our multiple linear regression equations or neural networks. There is also a need to devise better descriptor selection processes which can automatically generate an optimal subset of descriptors that results in

the most predictive model; such methods are under intense investigation [71–81] and are likely to result in improved QSAR/QPARs in the near future.

Acknowledgements

The authors would like to thank Revathy Iyer from the University of Pennsylvania for technical assistance. We would also like to acknowledge the National Textile Center (Grant No. C00-PH01) for financial support of this work.

References

- [1] Brown MA, Devito SC. Crit Revs Env Sci Technol 1993; 23:249.
- [2] Ashby J, Lefevre PA, Callander RD. Mut Res 1983; 116:271.
- [3] Garg A, Bhat KL, Bock CW. Dyes and Pigments 2002; 55:35.
- [4] Chung KT, Kirkovsky L, Kirkovsky A, Purcell WP. Mut Res 1997;387:1.
- [5] Abmann N, Emmrich M, Kampf G, Kaiser M. Mut Res 1997;395:139.
- [6] Rashid KA, Arjmand M, Sandermann H, Mumma ROJ. Environ Sci Health 1987;B22(6):721.
- [7] Mori Y, Niwa T, Toyoshi K, Hirano K, Sugiura M. Mut Res 1983;121:95.
- [8] Chung KT. Mut Res 1983;114:269.
- [9] Degawa M, Shoji Y, Masuko K, Hashimoto Y. Can Lett 1979;8:71.
- [10] Freeman HS, Esancy M, Esancy JF, Claxton JD. Chem Technol 1991:439.
- [11] CODESSA™, v2.0, Semichem, 7204 Mullen, Shawnee, KS 66216, USA.
- [12] AMPAC 5.0, ©1994 Semichem, 7128 Summit, Shawnee, KS 66216, USA.
- [13] Huibers PDT, Lobanov VS, Katritzky AR, Shah DO, Karelson M. Langmuir 1996;12:1462.
- [14] Katritzky AR, Sild S, Karelson M. J Chem Inf Comput Sci 1998;38:1171.
- [15] Katritzky AR, Karelson M, Lobanov VS. Pure Appl Chem 1997;69:245.
- [16] Mu L, Drago RS, Richardson DEJ. Chem Soc Perkin 1998;II:159.
- [17] Bhat KL, Freeman HS, Velga J, Sztandera L, Trachtman M, Bock CW. Dyes and Pigments 2000;46:109.
- [18] Bhat KL, Trachtman M, Bock CW. Dyes and Pigments 2001;48:197.
- [19] Bhat KL, Garg A, Trachtman M, Bock CW. Dyes and Pigments 2001;50:133.

- [20] SPARTAN v5.0, Wavefunction Inc., 18401 Von Karmen Avenue, Suite 370, Irvine, CA 92612, USA.
- [21] Bhat KL, Garg A, Bock CW. *Dyes and Pigments* 2002; 52:145.
- [22] Debnath AK, Debnath G, Shusterman AJ, Hansch C. *Env Mol Mutagen* 1992;19:37.
- [23] Zhang L, Sannes K, Shusterman AJ, Hansch C. *Chem Biol Interact* 1992;81:149.
- [24] ACD module: v4.5, Advanced Chemistry Development, Inc., 90 Adelaide St. W., Suite 702, Toronto, Ontario, Canada M5H 3V9.
- [25] Sztandera LM. *J Appl Comp Sci* 2001;9(1):43.
- [26] Sztandera LM. *J Artificial Neural Systems* 1994;1:41.
- [27] Sztandera LM. *Information Sciences Journal* 1995;3(2):75.
- [28] Cios KJ, Sztandera LM. *Neurocomputing* 1997;14:383.
- [29] Freeman HS, Esancy JF, Claxton JD. In: Peters AT, Freeman HS, editors. *Colour chemistry—the design and synthesis of organic dyes and pigments*. Elsevier Scientific Publications; 1991, Ch.4, p.85.
- [30] Hashimoto Y, Watanabe HK, Degawa M. *Gann* 1981; 72:921.
- [31] Miller JA, Miller EC. *Com Res* 1961;21:1068.
- [32] Sato K, Poirier LA, Miller JA, Miller EC. *Can Res* 1966; 26:1678.
- [33] Mori Y, Niwa T, Hori T, Toyoshi K. *Carcinogenesis* 1980; 1:121.
- [34] Degawa M, Kanazawa C, Hashimoto Y. *Carcinogenesis* 1982;3:1113.
- [35] Miller EC, Kadlubar FF, Miller JA, Pitot HC, Drinkwater NR. *Can Res* 1979;39:3411.
- [36] Mori Y, Hori T, Toyoshi K, Horie MJ. *Pharm Dyn* 1979; 1:192.
- [37] Miller JA, Sapp RW, Miller EC. *Can Res* 1949;9:652.
- [38] Degawa M, Miyairis, Hashimoto Y. *Gann* 1978;69:367.
- [39] Yahagi T, Degawa M, Seino Y, Matsushima T, Nagao M, Sugimura T, Hashimoto Y. *Can Letts* 1975;1:91.
- [40] Kitagawa T, Pitot HC, Miller EC, Miller JA. *Can Res* 1979;39:112.
- [41] Hashimoto Y, Watanabe H, Degawa M. *Gann* 1977; 68:373.
- [42] Degawa M, Hashimoto Y. *Chem Pharm Bull* 1976; 24:1485.
- [43] *Aldrich handbook of fine chemicals and laboratory equipment*, 2000–2001, p. 1159.
- [44] Lipinski CA, Lombardo F, Domiay BW, Freeney PJ. *Advanced Drug Delivery Reviews* 1997;23:3.
- [45] Rohrbach RH, Jurs PC. *Analytica Chimica Acta* 1987; 199:99.
- [46] Zefirov NS, Kirpichenok MA, Izmailov FF, Trofimov MI. *Dokl Akad Nauk SSSR* 1997;296:886.
- [47] Kirpichenok MA, Zefirov NS. *Zh Org Khim* 1987;23:4.
- [48] Stanton DT, Jurs PC. *Anal Chem* 1990;62:2323.
- [49] Stanton DT, Egolf LM, Jurs PC, Hicks MG. *J Chem Inf Comput Sci* 1992;32:306.
- [50] Fukui K. *Theory of orientation and stereoselection*. Berlin: Springer-Verlag; 1975.
- [51] Sanderson RT. *J Chem Ed* 1988;65:112.
- [52] Huheey JE, Keiter EA, Keiter RL. *Inorganic chemistry: principles of structure and reactivity*. 4th ed. New York, USA: Harper Collins; 1993.
- [53] Shannon CE. *Bell System Technical Journal* 1948;27:379.
- [54] Sloane NJA, Wyner AD, editors. *IEEE Press, Claude Elwood Shannon: Collected papers*, New York, 1993.
- [55] Personal communication from Holder A.
- [56] Hashimoto Y, Degawa M, Watanabe HK, Tada M. *Gann* 1981;72:937.
- [57] Lin JK, Schmall B, Sharpe ID, Miuva I, Miller JA, Miller EC. *Can Res* 1975;35:832.
- [58] Kabuldar FF, Miller JA, Miller EC. *Can Res* 1976; 36:2350.
- [59] Mager DE, Jusko WJ. *J Pharm Sci* 2002;91:2441.
- [60] Maran U, Karelson M, Katritzky AR. *Quant Struct-Acta Relat* 1999;18:3.
- [61] Irie B, Miyake S. *Proc IEEE Int Conf Neural Networks* 1988:641.
- [62] Balaban A. *Pure Appl Chem* 1983;55:199.
- [63] Schaffer C. *Machine Learning* 1993;13:135.
- [64] Benigni R, Andreoli C, Giuliani A. *Env Mol Mutagen* 1994;24:208.
- [65] Enslein K, Borgstedt HH. *Toxicol Lett* 1989;49:107.
- [66] Rosenkranz HS, Klopman G. *Mut Res* 1989;221:217.
- [67] Rose SL, Juys PC. *J Med Chem* 1982;25:769.
- [68] Claxton LD, Walsh DB, Esancy JF, Freeman HS. *Prog Clin Biol Res* 1990;340:11.
- [69] Stead CV. In: Shore J, editor. *Chemistry of azo colorants in colorants and auxiliaries*, vol. 1. Society of Dyers and Colorists; 1990. p. 147.
- [70] Dicesare N, Lakowicz JR. *Org Letts* 2001;3:3891.
- [71] Izrailev S, Agrafiotis DK. *SAR and QSAR in Environmental Research* 2002;13:417.
- [72] Izrailev S, Agrafiotis DK. *J Chem Inf Comput Sci* 2001; 41:176.
- [73] Dorigo M, Caro G, Gambardella LM. *Artif Life* 1999; 5:137.
- [74] Wanchana S, Yamashita F, Hashida M. *Pharmazie* 2002; 57:127.
- [75] Selwood DL, Livingstone DJ, Comley JCW, O'Dowd AH, Hudson AT, Jackson P, Jandu KS, Rose VS, Stables JN. *J Med Chem* 1990;33:136.
- [76] Agrafiotis DK, Cedeno W. *J Med Chem* 2002;45:1098.
- [77] Yasri A, Hartsough D. *J Chem Inf Comput Sci* 2001; 41:1218.
- [78] Burden FR, Ford MG, Whitley DC, Winkler DA. *J Chem Inf Comput Sci* 2000;40:1423.
- [79] Sutter JM, Dixon SL, Jurs PC. *J Chem Inf Comput Sci* 1995;35:77.
- [80] Luke BT. *J Chem Inf Comput Sci* 1994;34:1279.
- [81] Kubinyi H. *QSAR* 1994;13:285.