Carcinogenesis vol.22 no.9 pp.1561–1571, 2001

Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure–activity relationships model

Rainer Franke, Andreas Gruska, Alessandro Giuliani[1] and Romualdo Benigni[1,2]

Consulting in Drug Design GbR, Gartenstraβe 14, D-16352 Basdorf, Germany and [1]Istituto Superiore di Sanita', Laboratory of Comparative Toxicology and Ecotoxicology, Viale Regina Elena 299, I-00161 Rome, Italy

[2]To whom correspondence should be addressed
Email: rbenigni@iss.it

The aromatic amines are widely used industrial chemicals and can be found in tobacco smoke as well as in products generated during cooking. In a previous study, we established quantitative structure–activity relationship (QSAR) models linking the carcinogenic potency of non-heterocyclic carcinogenic aromatic amines to a series of molecular determinants. We also found that QSAR models for carcinogenic potency were inadequate in describing the difference between carcinogenic and non-carcinogenic amines [Benigni,R., Giuliani,A., Franke,R. and Gruska,A. (2000) *Chem. Rev.*, 100, 3697–3714]. In this paper, we derived specific QSAR models for separating active from inactive amines. It appeared that hydrophobicity (as measured by the octanol/water partition coefficient, log*P*) played a major role in modulating the potency of the carcinogens, whereas mainly electronic (reactivity) and steric characteristics separated the carcinogens from the non-carcinogens. Interestingly, a similar pattern was previously demonstrated by us regarding their mutagenic activity [Benigni,R., Passerini,L., Gallo,G., Giorgi,F. and Cotta-Ramusino,M. (1998) *Environ. Mol. Mutagen.*, 32, 75–83]. Based on the QSAR models found, the molecular determinants of the mechanisms of action of aromatic amines are discussed in detail. The QSAR models obtained can be used directly for estimating the carcinogenicity of other non-heterocyclic aromatic amines for which experimental data are not available. With the QSARs in Benigni *et al.* (2000) and the present results, a two-step prediction of carcinogenicity of aromatic amines is possible: (i) step 1, yes/no activity from the discriminant functions; and (ii) step 2, if the answer from step 1 is yes then prediction of the degree of potency from the equations in Benigni *et al.* (2000). Thus, QSAR models can contribute to the following: the direct synthesis of safer chemicals; the estimation of the risk posed by amines present in the environment; setting priorities for further experimentation, thus also reducing the use of experimental animals. Whereas the quality of *in vivo* experimental data is often questioned, the robustness and intepretability of the present results strongly support the reliability of the rodent carcinogenicity assay.

Abbreviations: BRM, carcinogenic potency in mice; BRR, carcinogenic potency in rats; QSAR, quantitative structure–activity relationship; SAR, structure–activity relationship.

Introduction

The class of aromatic amines has a special place among the chemical carcinogens, in terms of extent of use and available scientific knowledge. Exposure to aromatic amines occurs in different industrial and agricultural activities as well as in tobacco smoking. Aromatic amines have been used as antioxidants in the production of rubber and in cutting oils, as intermediates in azo dye manufacturing and as pesticides. They are a common contaminant in several working environments, including the chemical and mechanical industries, and arylamines based dyes are widely used in the textile industry and in cosmetics (3). The wide use of aromatic amines together with the presence of relatively specific, very high exposures permitted the development of epidemiological knowledge unparalleled in other chemical classes. Bladder cancer in men is the most studied tumour type: a large number of studies relating professional exposure to arylamines and bladder cancer have been published (for a comprehensive review, see ref. 3). The odds ratios for arylamine exposure go from around 2 (i.e. a 2-fold increase in the probability of developing bladder cancer) for very mild exposures, increasing to around 100 for extremely high exposures (3–6).

The level of use and industrial importance of the aromatic amines have stimulated an enormous amount of investigations, both epidemiological and experimental (3,7–9). The molecular determinants of the mechanisms of action of the aromatic amines have also been studied with methods based on structure–activity relationship (SAR) and quantitative structure–activity relationship (QSAR) concepts, mostly regarding their mutagenic properties (1). We have recently presented the first detailed QSAR analysis of the molecular determinants that rule the carcinogenic potency gradation within the subset of carcinogenic non-heterocyclic aromatic amines (1).

The following QSAR models were derived for the gradation of carcinogenic potency in carcinogenic non-heterocyclic aromatic amines (BRM, carcinogenic potency in mice; BRR, carcinogenic potency in rats)

$$BRM = 0.88(\pm 0.27)\ \log P^{*} I(\mathrm{monoNH_2}) + 0.29(\pm 0.20)\ \log P^{*} I(\mathrm{diNH_2}) \quad (1)$$

$$+ 1.38(\pm 0.76)\ EHOMO - 1.28(\pm 0.54)\ ELUMO - 1.06(\pm 0.34)\ \Sigma MR_{2,6}$$

$$- 1.10(\pm 0.80)\ MR_3 - 0.20(\pm 0.16)\ E_S(R) + 0.75(\pm 0.75)\ I(\mathrm{diNH_2})$$

$$+ 11.16(\pm 6.68)$$

$$n = 37,\ r = 0.907,\ r^2 = 0.823,\ s = 0.381,\ F = 16.3,\ P < 0.001$$

$$BRR = 0.35(\pm 0.18)\ \log P + 1.93(\pm 0.48)\ I(\mathrm{Bi}) + 1.15(\pm 0.60)\ I(F) - 1.06(\pm 0.53)\ I(\mathrm{BiBr}) \quad (2)$$

$$+ 2.75(\pm 0.64)\ I(\text{RNNO}) - 0.48(\pm 0.30)$$

$$n = 41,\ r = 0.933,\ r^2 = 0.871,\ s = 0.398,\ F = 47.4,\ P < 0.001$$

where $BRM = \log (MW/TD_{50})_{mouse}$ and $BRR = \log (MW/TD_{50})_{rat}$. The TD_{50} is the daily dose rate required to halve the probability of an experimental animal remaining tumourless to the end of its standard life-span (10). MW is the molecular weight. The physical chemical and structural parameters in the equations are as follows: log*P*, logarithm of the partition coefficient in the system *n*-octanol/water; *EHOMO*, energy of the highest occupied molecular orbital; *ELUMO*, energy of the lowest unoccupied molecular orbital; $\Sigma MR_{2,6}$, sum of molar refractivity of substituents in the ortho-positions of the aniline ring; MR_3, molar refractivity of substituents in the meta-position of the aniline ring; $E_S(R)$, Charton's substituent constant for substituents at the functional amino group; *I*(monoNH2) = 1 for compounds with only one amino group; *I*(diNH2) = 1 for compounds with more than one amino group; *I*(Bi) = 1 for biphenyls; *I*(BiBr) = 1 for biphenyls with a bridge between the phenyl rings; *I*(RNNO) = 1 for compounds with the group N(Me)NO.

The terms in the equations point to the physical chemical determinants that govern the carcinogenic potency gradation, whereas the signs (+ and –) indicate the direction of the effects (increasing or decreasing). The key factor for carcinogenic potency is hydrophobicity, expressed by log*P*. Both *BRM* and *BRR* increase with increasing hydrophobicity. In the case of *BRM* the influence of hydrophobicity is stronger for compounds with one amino group [characterized by the indicator variable *I*(monoNH2)] in comparison with compounds with more than one amino group [characterized by the indicator variable *I*(diNH2)] (see the different coefficients 0.88 and 0.29). For *BRM*, electronic factors also play a role: potency increases with increasing *EHOMO* and with decreasing *ELUMO*. Such effects seem to be less important for BRR: no electronic terms occur in Equation 2. Carcinogenic potency also depends on the type of the ring system: aminobiphenyls [indicator variable *I*(Bi)] and, in the case of *BRR*, also fluorenamines [indicator variable *I*(*F*)] are intrinsically more active than anilines or naphthylamines. A bridge between the rings of the biphenyls decreases potency [*I*(BiBr)]. Steric factors are involved in the case of *BRM* but cannot be detected in the case of *BRR*. *BRM* strongly decreases with bulk in the positions adjacent to the functional amino group, and bulky substituents at the nitrogen and in position 3 also decrease potency. The latter effects are, however, not so important. In the case of *BRR*, *R* = (Me)NO strongly enhances potency (compounds with this substituent have no measured value for *BRM*). [The variables *I*(monoNH2), *I*(diNH2), *I*(Bi), *I*(F), *I*(BiBr) and *I*(RNNO) are called indicator variables, and indicate the presence (value = 1) or absence (value = 0) of the feature that they are coding for. For example, the presence of only one NH_2 group in the molecule implies that $I(\text{monoNH}_2) = 1$ and $I(\text{diNH}_2) = 0$; thus only the coefficient of $I(\text{monoNH}_2)$ (0.88) contributes to the value taken by *BRM*.]

The two equations model the gradation of potency of the carcinogenic aromatic amines, but not all aromatic amines are carcinogenic. Even though Equations 1 and 2 have high descriptive power for the gradation of potency, they did not describe inactive compounds correctly: in fact, we found that the non-carcinogenic amines were predicted to be weakly active by applying the two equations (1). In other terms, the models did not discriminate between weakly active amines and inactive amines. This indicates that yes/no activity depends, to some extent, on different molecular properties than the gradation of carcinogenic potency within the active compounds; this has been already shown for the mutagenic activity and potency of aromatic amines by Benigni *et al.* (2).

In the present work discriminant functions are generated which can separate carcinogenic from non-carcinogenic aromatic amines. The derived discriminant functions point to the molecular determinants that make the difference between the two sets of aromatic amines, thus lending themselves to scientific rationalization. In addition the present and previous QSAR models, combined together, provide a reliable tool for estimating the carcinogenicity of yet untested aromatic amines. Predictions can now be made in two steps: (i) yes/no prediction with the help of these discriminant functions; and (ii) for compounds predicted to be carcinogenic, prediction of the degree of carcinogenic potency from the QSARs presented in Benigni *et al.* (1).

Materials and methods

Chemical structures and parameters

Table I summarizes the compound structures (anilines, biphenylamines, naphthylamines and aminofluorenes) for which carcinogenicity data were available. Chemical structures are presented as substituted anilines according to the conventions outlined below. Table I also reports the biological activity information (see below). Table II gives the running number, CAS and common names of the chemicals. Not included in the analysis are compound nos 67 and 68, because of the highly reactive substituent at the amino group, and compound no. 11 because of its very special structure.

Structures were optimized by AM1 and the energies of the highest occupied (*EHOMO*) and the lowest unoccupied molecular orbital (*ELUMO*) were calculated (program system SYBYL; Tripos). Hydrophobicity is expressed in terms of log*P* (logarithm of the octanol/water partition coefficient) computed from the program TSAR (Oxford Molecular, Oxford, UK). These global parameters are presented in Table III.

Steric properties of ring substituents are characterized by molar refractivity (MR; values multiplied by 10^{-1}). For this purpose positions need to be defined. This was done in the same way as in Benigni *et al.* (1) with the following conventions: (i) the functional amino group is always in position 1, additional amino groups are treated as substituents; (ii) if more than one amino group is present, the one that is considered to be the functional group is that which has a substituent in an adjacent position (ortho-substituent); (iii) if only one ortho-substituent is present, this substituent is placed in position 2. Biphenyls, naphthalenes and fluorenes were treated as substituted anilines. For the biphenyls (Figure 1) substituents in the aniline part are characterized as in substituted anilines. In cases 1 and 2, the second part of the molecule (second phenyl ring plus substituents at this ring) is then treated as a para-substituent where bridge *x* may be present or absent. In case 3, the non-aniline part appears as ortho-substituent [there is only one compound of this type (no. 9)]. In the case of the naphthalenes (Figure 2) two situations are possible. They are treated as anilines substituted by $-C_4H_4-$ with an estimated *MR* of 1.6. This amount is equally distributed over the positions of substitution so that $MR_2 = MR_3 = 0.8$ in case 1, and $MR_3 = MR_4 = 0.8$ in case 2. If additional substituents occur the MR-values are correspondingly corrected. For the fluorenes, finally, steric effects are parameterized following the scheme presented in Figure 3. Charton's E_S values used in Benigni *et al.* (1) to characterize substituents at the amino nitrogen did not lead to satisfactory results in discriminant analysis. Therefore, the Sterimol parameters B_5 (maximal width) and *L* (length) were applied. Since these parameters are not additive in N-disubstituted compounds their use is restricted to monosubstituted derivatives. As a consequence, compound nos 8, 22, 23, 45, 46, 52 and 73 are not described by discriminant functions containing sterimol parameters for substituents at the amino nitrogen. In order to reintroduce disubstituted compounds into the analyses N-substituents are also characterized by molar refractivity (taking the sum in N-disubstituted compounds) and by indicator variables (see below). All steric parameters are summarized in Table IV.

Some of the special features of the compounds are characterized by the following indicator variables: *I*(BiBr) = 1 for biphenyls with a bridge between

the phenyl rings; $I(\mathrm{An}) = 1$ for anilines; $I(\mathrm{NO_2}) = 1$, if a NO_2 group is present; $I(\mathrm{monoNH_2}) = 1$, if only one amino group is present; $I(\mathrm{diNH_2}) = 1$, if more than one amino group is present; $I(\mathrm{NR}) = 1$, if the amino nitrogen is substituted; $I(\mathrm{o\text{-}NH_2}) = 1$, if a non-substituted amino group occurs in ortho-position to the functional amino group (also non-substituted).

Biological data

The rodent carcinogenicity data were derived from two sources. The primary source was the results of the National Cancer Institute/National Toxicology Program (NCI/NTP) carcinogenicity experiments, for which we considered the Selkirk and Soward's compilation (11), and for the more recent experiments, the individual NTP Technical Reports. The updated NTP database is also available on the internet at http://ntp-server.niehs.nih.gov/htdocs/pub.html. Only for chemicals not included in the NCI/NTP database, did we use the carcinogenicity database established by Gold *et al*. (10). This database is also available at http://potency.berkeley.edu/cpdb.html. In the case of chemicals present in both databases and tested in more than one laboratory, we used the NTP data. This was because of the well controlled and constant protocols used by the NTP (11).

Out of the 82 chemicals listed in Table I, three were not considered in the analyses for structural reasons (see above). Out of the 79 chemicals actually considered, the majority (55) were derived from the NTP database and 24 from Gold's database. The chemicals tested by NTP were (numbered in Table I) 7, 9, 10, 12–53, 55–58, 73, 74, 76 and 79–81.

All biological data are summarized in Table I, which presents carcinogenicity scores for the four rodent experimental groups (rat, mouse, male, female) and for overall carcinogenicity. The scoring for the experimental groups is from 1 to 4 ('no evidence', 'equivocal evidence', 'some evidence', 'clear evidence'); for the overall value it is from 1 to 3 ('inactive', 'equivocal' or 'borderline', 'active'). In the overall carcinogenicity scores, we relied on the traditional NCI/NTP classification, that considers any chemical as a carcinogen that has at least 'some evidence' of carcinogenicity in one experimental system. The scoring of 1–4 for the experimental groups follows the recent classification by NCI/NTP. For older experiments, the carcinogens (no distinction between 'some evidence' and 'clear evidence') were given a score of 4. These scores are only ordinal scores and were not used as quantitative values in the mathematical analyses. Moreover, in order to construct the QSAR models only on the most reliable results, the 'equivocal evidence' chemicals were discarded. The 'some evidence' carcinogens form too small a class, so they were not considered in the discriminant analyses of the individual experimental groups. Therefore, the carcinogenicity classes were then defined as follows: (i) overall carcinogenicity: class 1 (inactive compounds), score 1; class 2 (active compounds), score 3; (ii) rat and mice carcinogenicity: class 1 (inactive compounds), score 1; class 2 (active compounds), score 4.

Discriminant analysis

Non-elementary discriminant analysis (12) was used resulting in canonical discriminant functions of the general form

$$w = \Sigma\, a_i x_i \tag{3}$$

The x_i are molecule parameters related to the distribution of compounds over the classes, and the coefficients a_i are so determined that the separation of classes is optimal. For a two-class case the discriminant function describes a one-dimensional coordinate system with w as axis. On this axis, the two classes have different positions characterized by their means of w (Figure 4). A compound is assigned to that class whose mean is closer to its computed value of w. In a case such as in Figure 4 [$w_{(\mathrm{mean,class2})} < w_{(\mathrm{mean,class1})}$] all negative terms in the discriminant function support the assignment to this class. The reverse is, of course, true if $w_{(\mathrm{mean,class2})} > w_{(\mathrm{mean,class1})}$. The mutual position of classes on the w axis does not depend on which of the two classes contains the active compounds; therefore, the class means of w are given for each discriminant function as, otherwise, an interpretation of the discriminant function would not be possible. The discriminant functions are presented in their standardized form as this reflects more clearly the relative weight of the variables. Thus, in order to compute w values from these functions, the variables need to be standardized. To check the validity of a discriminant function the techniques of reclassification and of cross validation are used. In reclassification the compounds are assigned to a class with the help of w values computed from the discriminant function. A discriminant function with a high descriptive power then shows a high correct reclassification rate. Cross validation is a simulated prediction. Groups of compounds are left out from the analysis, and discriminant functions are then evaluated from the remaining compounds. The class membership of the compounds left out is then computed from the respective discriminant function. In the present paper three cross validation groups, with a third of the data deleted in each group, were used. Functions with good predictive power should show a high rate of correct predictions (the correct rate from cross validation is, in most cases, somewhat lower than the reclassification rate).

Results

The rodent carcinogenicity data listed in Table I were analyzed with discriminant analysis. With this mathematical classification method we constructed equations that, based on the chemical descriptors of the aromatic amines, separated the carcinogens from the non-carcinogens. The presence and the relative importance of the chemical descriptors in the equations are discussed in terms of the aromatic amines' mechanisms of action. We applied this procedure separately to the overall carcinogenicity classification, as well as to the four rodent experimental groups (rat, mouse, male, female). To obtain the most reliable QSAR models, we discarded the 'equivocal evidence' results. The 'some evidence' results were too few to be considered as a separate class, and were not included in the analyses of the individual experimental groups. The classes used in discriminant analysis were Class 1, non-carcinogens (correspondent to Score 1 in Table I) and Class 2, carcinogens (correspondent to Score 4 for rat and mouse carcinogenicity, and Score 3 for overall carcinogenicity in Table I). All discriminant functions presented are statistically significant at $P < 0.01$.

Overall carcinogenicity

The discriminant function

$$\begin{aligned} w = &-2.86\, L(R) + 2.65\, B_5(R) - 1.16\, EHOMO + \\ &1.76\, ELUMO + 0.40\, MR_3 \\ &+ 0.58\, MR_5 + 0.54\, MR_6 - 1.55\, I(\mathrm{An}) + 0.74\, I(\mathrm{NO_2}) \\ &- 0.55\, I(\mathrm{BiBr}) \end{aligned} \tag{4}$$

$w_{(\mathrm{mean,class1})} = -1.56 \quad N_1 = 13$

$w_{(\mathrm{mean,class2})} = 0.38 \quad N_2 = 53$

where N_1 = number of non-carcinogens (Class 1) and N_2 = number of carcinogens (Class 2), correctly reclassifies 87.9% of the compounds (Class 1, 84.6%; Class 2, 88.7%) and is stable in cross validation (correct classification: all compounds = 84.8%; Class 1, 84.6%; Class 2, 84.8%). Eight compounds were not correctly reclassified (nos 16, 24, 35, 37, 50, 55, 69 and 71). If only two of them, compounds nos 16 and 50, were removed from the analysis, the correct reclassification rate and correct cross validation become 95.3% (Class 1, 100%; Class 2, 94.2%; misclassified compounds: nos 37, 55 and 71) with the same variables appearing in the discriminant function

$$\begin{aligned} w = &\ 3.42\, L(R) - 3.11\, B_5(R) + 1.57\, EHOMO \\ &-2.19\, ELUMO - 0.66\, MR_3 \\ &- 0.65\, MR_5 - 0.54\, MR_6 + 1.64\, I(\mathrm{An}) - 0.57\, I(\mathrm{NO_2}) + 0.63 \\ &I(\mathrm{BiBr}) \end{aligned} \tag{5}$$

$w_{(\mathrm{mean,class1})} = 2.04 \quad N_1 = 12$

$w_{(\mathrm{mean,class2})} = -0.47 \quad N_2 = 52$

The classes are well separated by the discriminant functions which show high correct reclassification rates and stability in cross validation.

In contrast to the QSAR equations describing carcinogenic potency within the active compounds (Equations 1 and 2), no log$P$ term appears in these functions, so that hydrophobicity does not appear to be a key factor in class separation. There is, however, a significant multiple correlation between log$P$ and *EHOMO*, *ELUMO*, MR_6 and I(BiBr) with $r = 0.73$, so that some hydrophobic effect might be hidden behind these parameters. The probability of a compound to be assigned to the active class increases with decreasing values of *EHOMO* and increasing values of *ELUMO*, the contribution of *ELUMO*

Table I. Chemical structures, CAS numbers and carcinogenic activity

No.	CAS	Structure				Carcinogenic activity				
		Ring	Bridge	AnX	R	Overall	Male rats	Female rats	Male mice	Female mice
1	28322-02-3	F		2,3-Me2	H, COMe	–	–	1	–	–
2	91-59-8	N		3-C4H4-4	H2	3	4	4	4	4
3	92-87-5	Bi		4-(Ph-4-NH2)	H2	3	4	4	4	4
4	53-96-3	F		3,4-Me2	H, COMe	3	4	4	4	4
5	101-21-3	An		3-C1	H, COOiPr	1	1	1	1	1
6	101-14-4	Bi	CH2	2-C1,4(CH2-Ph-3-C1,4-NH2)	H2	3	4	4	–	4
7	636-21-5	An		2-Me	H2	3	4	4	4	4
8	2465-27-2	Bi	C=NH2	4-(C(=NH)-Ph-4-N(Me)2)	Me2	3	4	–	4	4
9	2185-92-4	Bi		2-Ph	H2	3	1	1	2	4
10	609-20-1	An		2,6-C12,4-NH2	H2	3	1	1	4	4
11	2784-94-3	An		2-NO2,4-N(C2H4OH)2	H, Me	3	2	3	4	4
12	13552-44-8	Bi	CH2	4-(CH2-Ph-4-NH2)	H2	3	4	4	4	4
13	150-68-5	An		4-C1	H, CONMe2	3	4	1	1	1
14	101-80-4	Bi	O	4-(O-Ph-4-NH2)	H2	3	4	4	4	4
15	87-62-7	An		2,6-Me2	H2	3	4	4	–	–
16	15481-70-6	An		2-Me,3-NH2	H2	1	1	1	1	1
17	20265-97-8	An		4-OMe	H2	2	2	1	1	1
18	118-92-3	An		2-COOH	H2	1	1	1	1	1
19	140-49-8	An		4-COCH2C1	H, COMe	1	1	1	1	1
20	61702-44-1	An		2-C1,4-NH2	H2	1	1	1	1	1
21	54150-69-5	An		2,4-OMe2	H2	1	1	1	1	1
22	315-18-4	An		2,6-Me2,4-OCONMe	Me2	1	1	1	1	1
23	1465-25-4	N		2-C4H4-3	H, C2H4NH2	1	1	1	1	1
24	619-17-0	An		2-COOH,5-NO2	H2	1	1	1	1	1
25	99-56-9	An		2-NH2,4-NO2	H2	1	1	1	1	1
26	624-18-0	An		4-NH2	H2	1	1	1	1	1
27	101-54-2	Bi	NH	4-NH-Ph	H2	1	1	1	1	1
28	103-85-5	An		H	H, CSNH2	1	1	1	1	1
29	6369-59-1	An		2-Me,4-NH2	H2	1	1	1	1	1
30	17026-81-2	An		2-OEt,5-NHCOMe	H2	3	1	1	4	1
31	132-32-1	F		3-Me,4-NEt	H2	3	4	4	4	4
32	119-34-6	An		3-NO2,4-OH	H2	3	4	2	1	1
33	142-04-1	An		H	H2	3	4	4	1	1
34	134-29-0	An		2-OMe	H2	3	4	4	4	4
35	20265-96-7	An		4-C1	H2	3	4	2	3	1
36	5131-60-2	An		2-C1,5-NH2	H2	3	4	1	1	4
37	95-83-0	An		2-NH2,4-C1	H2	3	4	4	4	4
38	95-74-9	An		2-C1,4-Me	H2	1	1	1	1	1
39	102-50-1	An		2-Me,4-OMe	H2	3	4	4	–	1
40	120-71-8	An		2-OMe,5-Me	H2	3	4	4	4	4
41	80-08-0	Bi	SO2	4-(SO2-Ph-4-NH2)	H2	3	4	1	1	1
42	615-05-4	An		2-OMe,5-NH2	H2	3	4	4	4	4
43	950-80-7	An		2-Me,5-NH2	H2	3	4	4	1	4
44	2164-17-2	An		3-CF3	H, CONMe2	2	1	1	2	1
45	101-61-1	Bi	CH2	4-(CH2-Ph-4-N(Me)2)	Me2	3	4	4	2	4
46	90-94-8	Bi	CO	4-(CO-Ph-4-N(Me)2)	Me2	3	4	4	4	4
47	2243-62-1	N		2-C3H3C(NH2)-3	H2	3	1	4	4	4
48	1777-84-0	An		3-NO2,4-OEt	H, COMe	3	1	1	4	1
49	99-59-2	An		2-OMe,5-NO2	H2	3	1	1	4	4
50	5307-14-2	An		2-NO2,4-NH2	H2	3	1	1	1	4
51	139-65-1	Bi	S	4-(S-Ph-4-NH2)	H2	3	4	4	4	4
52	1582-09-8	An		2,6-(NO2)2,4-CF3	(nPr)2	3	1	1	1	4
53	137-17-7	An		2,4,5-Me3	H2	3	4	4	2	4
54	92-67-1	Bi		4-Ph	H2	3	–	–	4	4
55	121-88-0	An		2-OH,4-NO2	H2	3	3	1	1	1
56	135-88-6	N		3-C4H4-4	H, Ph	2	1	1	1	2
57	137-09-7	An		2-OH,5-NH2	H2	3	1	1	3	1
58	100-01-6	An		4-NO2	H2	2	–	–	2	1
59	4075-79-0	Bi		4-Ph	H, COMe	3	–	4	–	–
60	324-93-6	Bi		4-Ph-4-F	H2	3	–	–	4	4
61	398-32-3	Bi		4-Ph-4-F	H, COMe	3	4	–	–	–
62	363-17-7	F		3,4-Me2	H, COCF3	3	–	4	–	–
63	101-79-1	Bi	O	4-(OPh-4-C1)	H2	3	4	–	1	4
64	91-94-1	Bi		2-C1,4-(Ph-3-C1,4-NH2)	H2	3	4	4	–	–
65	77-46-3	Bi	SO2	4-SO2-Ph-4-NHCOMe	H, COMe	3	–	4	–	–
66	62-44-2	An		4-OEt	H, COMe	3	4	4	4	4
67	937-25-7	An		4-F	Me, NO	3	4	–	–	–
68	614-00-6	An		H	Me, NO	3	4	4	–	–

Table I. *continued*

No.	CAS	Structure				Carcinogenic activity				
		Ring	Bridge	AnX	R	Overall	Male rats	Female rats	Male mice	Female mice
69	615-28-1	An		2-NH2	H2	3	4	–	4	4
70	7411-49-6	Bi		2-NH2,4-(Ph-3,4-(NH2)2)	H2	3	4	–	4	1
71	63886-77-1	An		2,4,5,6-F4,3-NH2	H2	3	1	–	4	1
72	6334-11-8	An		2,4,6-Me3	H2	3	4	–	4	4
73	121-69-7	An		H	Me2	3	3	1	1	2
74	2871-01-4	An		3-NO2,4-NHC2H4OH	H2	2	1	1	2	–
75	540-23-8	An		4-Me	H2	3	1	–	4	4
76	99.57-0	An		2-OH,5-NO2	H2	3	3	1	1	1
77	634-93-5	An		2,4,6-C13	H2	3	1	–	4	1
78	638-03-9	An		3-Me	H2	3	1	–	4	2
79	20325-40-0	Bi		2-OMe,4-(Ph-3-OMe,4-NH2)	H2	3	4	4	–	–
80	612-82-8	Bi		2-Me,4-(pH-3-Me,4-NH2)	H2	3	4	4	–	–
81	133-90-4	An		2,5,C12,3-COOH	H2	3	1	1	2	4
82	838-88-0	Bi	CH2	2-Me,4-(CH2Ph-3-Me,4-NH2)	3	4	4	–	–	

Ring: An, aniline; Bi, biphenyl; N, naphthalene; F, fluorene. Bridge, bridge between the phenyl rings in biphenyls if such a bridge occurs. AnX, ring substituent with an aniline moiety as core according to the conventions presented in 2.1.; R, substituents at the nitrogen of the functional amino group.

being more important. Clearly, this effect it is opposite to that seen in Equation 1. This is also true for bulk in the meta-position: according to Equation 1, bulk in this position decreases carcinogenic potency of active amines while, according to the discriminant functions in Equations 4 and 5, bulk in the meta-positions supports the assignment of compounds to the active class. As explained in the Discussion, these 'opposite' contributions in the two models are likely to point to the existence of optimal values of the parameters for maximum activity.

With respect to the ortho position, the probability of a compound becoming carcinogenic increases with MR_6. According to the conventions used, the appearance of a substituent in position 6 implies 2,6-disubstitution. There are four compounds of this type in the data set with either 2,6-Me_2 or 2,6-Cl_2 (nos 10, 15, 72 and 77) which all belong to the active class. The real meaning of the MR_6 term is that this substitution pattern obviously supports carcinogenic activity, which is in keeping with the statement 'substitution of a chloro group or methyl group or methoxy group ortho to the amino group often enhances potency' (7). Again, the situation for the gradation of potency (Equation 1) is different where bulk in this position appears to be unfavorable for the degree of carcinogenic potency in mice. $R(L)$ and $R(B_5)$ in Equations 4 and 5 cannot be replaced by $MR(R)$. This combination is not easy to interpret as both variables are highly correlated ($r = 0.97$; this is the only disturbing correlation in the data set). $R(B_5)$ can be replaced by $R(L)^2$ indicating a possible optimum in substituent length at $L(R)_0 \approx 3.1$. This optimum is close to the length of COMe [$L(R) = 4.06$]. A possible explanation for the steric terms for the amino substituents then is that bulk in this position is unfavorable except for the COMe substituent. Obviously, the key factors governing yes/no activity and the gradation of potency within the active compounds are different; for a more detailed discussion of this point, see Discussion. Finally, other things being equal, the probability of being active decreases if a compound belongs to the group of anilines and if a bridge between the two rings in the biphenyl compounds occurs and increases if NO_2 groups are present.

Rat carcinogenicity

Compound nos 50 and 65 behave as outliers and drastically reduce the sharpness of class separation. For this reason they were eliminated. The following discriminant function achieves a highly significant separation of classes for female rat carcinogenicity (without nos 50 and 65)

$$w = 0.65\,L(R) + 0.79\,EHOMO - 1.54\,ELUMO + 0.76\,MR_2 - 0.50\,MR_5 + 1.32\,I(\mathrm{An}) - 0.53\,I(\text{o-NH2}) + 0.99\,I(\mathrm{BiBr}) + 0.99\,I(\mathrm{diNH_2}) - 1.08\,\log P^{*}I(\mathrm{diNH2}) \quad (6)$$

$w_{(\mathrm{mean,class1})} = 1.05 \quad N_1 = 30$

$w_{(\mathrm{mean,class2})} = -1.21 \quad N_2 = 26$

The correct reclassification rate of discriminant function (6) amounts to 91.1% (Class 1, 93.3%; Class 2, 88.5%; misclassified compounds: nos 2, 16, 41, 47 and 66) with a fairly stable cross validation (all compounds: 80.4%; Class 1, 76.7%; Class 2, 84.6%). If the omitted compound nos 50 and 65 are included the discriminant function remains stable but the correct reclassification rate is reduced to 84.5%.

Substitution at the amino nitrogen can also be described by a simple indicator variable instead of Verloop parameters without loss of separating power now including di-substituted derivatives (without compound nos 50 and 65):

$$w = 0.63\,I(\mathrm{NR}) + 1.13\,EHOMO - 1.53\,ELUMO + 0.65\,MR_2 - 0.38\,MR_5 + 1.24\,I(\mathrm{An}) - 0.40\,I(\text{o-}\mathrm{NH_2}) + 0.84\,I(\mathrm{BiBr}) + 0.73\,I(\mathrm{diNH_2}) - 1.10\,\log P^{*}I(\mathrm{diNH_2}) \quad (7)$$

$w_{(\mathrm{mean,class1})} = 0.94 \quad N_1 = 34$

$w_{(\mathrm{mean,class2})} = -1.11 \quad N_2 = 29$

Exchanging $L(R)$ against $I(NR)$ has no effect on the other terms of the discriminant function and leads to almost the same correct reclassification (all compounds: 94.1%; Class 1, 94.1%; Class 2, 86.2%), but cross validation shows a poorer result (all compounds: 73.0%; Class 1, 70.6%; Class 2, 75.9%). If $MR(R)$ is used instead of $I(NR)$, two more compounds (one from each class) are misclassified (correct reclassification rate 87.1%).

For male rat carcinogenicity a good separation of classes is achieved by the discriminant function in Equation 8 (without compound nos 32 and 78):

Table II. Running number, CAS and common names of the chemicals

No.	Chemical	CAS
1	4-Acetylaminofluorene	28322-02-3
2	2-Naphthylamine	91-59-8
3	Benzidine	92-87-5
4	2-Acetylaminofluorene	53-96-3
5	Isopropyl-*N*-(3-chlorophenyl)carbamate	101-21-3
6	4,4′-Methylenebis(2-chloroaniline)	101-14-4
7	*o*-Toluidine · HCl	636-21-5
8	Auramine	2465-27-2
9	2-Biphenylamine · HCl	2185-92-4
10	2,6-Dichloro-*p*-phenylenediamine	609-20-1
11	HC Blue N 1	2784-94-3
12	4,4′-Methylenedianiline · 2HCl	13552-44-8
13	Monuron	150-68-5
14	4,4′-Oxydianiline	101-80-4
15	2,6-Xylidine	87-62-7
16	2,6-Toluenediamine · 2HCl	15481-70-6
17	*p*-Anisidine · HCl	20265-97-8
18	*o*-Anthranilic acid	118-92-3
19	4-(Chloroacetyl)-acetylanilide	140-49-8
20	2-Chloro-*p*-phenylenediamine sulphate	61702-44-1
21	2,4-Dimethoxy analine · HCl	54150-69-5
22	Mexarbate	315-18-4
23	*N*-(1-Naphthyl)-ethylenediamine · 2HCl	1465-25-4
24	4-Nitroanthranilic acid	619-17-0
25	4-Nitro-*o*-phenylenediamine	99-56-9
26	*p*-Phenylenediamine · 2HCl	624-18-0
27	*N*-Phenyl-*p*-phenylenediamine	101-54-2
28	1-Phenyl-2-thiourea	103-85-5
29	2,5-Tolueneidamine sulfate	6369-59-1
30	3-Amino-4-ethoxyacetanilide	17026-81-2
31	3-Amino-9-ethylcarbazole · HCl	132-32-1
32	4-Amino-2-nitrophenol	119-34-6
33	Aniline · HCl	142-04-1
34	*o*-Anisidine · HCl	134-29-0
35	*p*-Chloroaniline · HCl	20265-96-7
36	4-Chloro-*m*-phenylenediamine	5131-60-2
37	4-Chloro-*o*-phenylenediamine	95-83-0
38	3-Chloro-*p*-toluidine	95-74-9
39	*m*-Cresidine	102-50-1
40	*p*-Cresidine	120-71-8
41	Dapsone(4,4-sulfonyldianiline)	80-08-0
42	2,4-Diaminoanisole sulfate	615-05-4
43	2,4-Diaminotoluene	950-80-7
44	Fluometuron	2164-17-2
45	4,4′-Methylene-bis-*N,N′*-dimethylaniline	101-61-1
46	Michler's ketone	90-94-8
47	1,5-Naphthalenediamine	2243-62-1
48	3-Nitro-*p*-acetophenetide	1777-84-0
49	5-Nitro-*o*-anisidine	99-59-2
50	2-Nitro-*p*-phenylenediamine	5307-14-2
51	4,4′-Thiodianiline	139-65-1
52	Trifuralin	1582-09-8
53	2,4,5-Trimethylaniline	137-17-7
54	4-Aminobiphenyl	92-67-1
55	2-Amino-5-nitrophenol	121-88-0
56	*N*-Phenyl-2-naphthylamine	135-88-6
57	2,4-Diaminophenol · 2HCl	137-09-7
58	*p*-Nitroaniline	100-01-6
59	4-Acetylaminobiphenyl	4075-79-0
60	4′-Fluoro-4-aminodiphenyl	324-93-6
61	*N*-(4-(4′-Fluorobiphenyl))acetamide	398-32-3
62	*N*-(2-Fluorenyl)-2,2,2-trifluoroacetamide	363-17-7
63	4-Chloro-4′-amino-diphenylether	101-79-1
64	3,3′-Dichlorobenzidine	91-94-1
65	4,4′-Sulphonylbisacetanilide	77-46-3
66	Phenacetin	62-44-2
67	*N*-Nitroso-*N*-methyl-4-fluoroaniline	937-25-7
68	Nitroso-*N*-Methylaniline	614-00-6
69	*o*-Phenylenediamine · 2HCl	615-28-1
70	3,3′,4,4′-Tetraaminobiphenyl · 4HCl	7411-49-6
71	Tetrafluoro-*m*-phenylenediamine · 2HCl	63886-77-1
72	2,4,6-Trimethylalinine · HCl	6334-11-8
73	N,*N*-Dimethylaniline	121-69-7
74	H.C. Red N 3	2871-01-4
75	*p*-Toluidine · HCl	540-23-8
76	2-Amino-4-nitrophenol	99-57-0
77	2,4,6-Trichloroaniline	634-93-5
78	*m*-Toluidine · HCl	638-03-9
79	3,3′-Dimethoxybenzidine · 2HCl	20325-40-0
80	3,3′-Dimethylbenzidine	612-82-8
81	Chloramben	133-90-4
82	4,4′-Methylenebis (2-methylaniline)	838-88-0

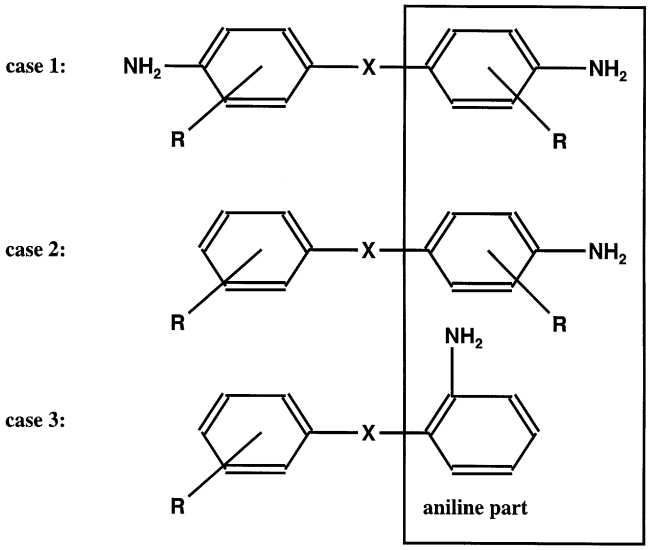

Fig. 1. Treatment of biphenyls.

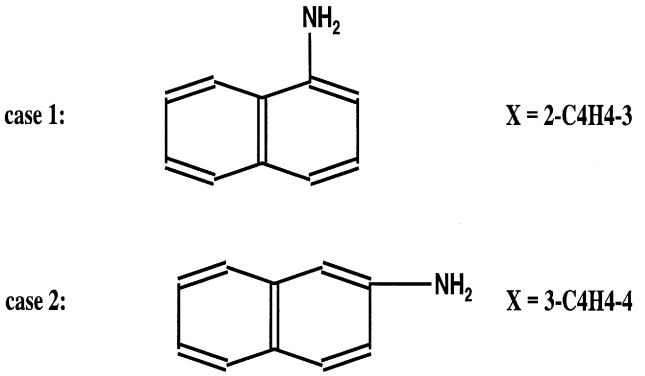

Fig. 2. Treatment of naphthalenes.

$$w = 0.48\,L(R) + 0.90\,EHOMO - 1.43\,ELUMO + 0.72\,MR_2 + 1.13\,I(\mathrm{An}) - 0.54\,I(\mathrm{o\text{-}NH_2}) - 0.45\,MR_5 + 0.70\,I(\mathrm{diNH_2}) - 0.80\,\log P^{*} I(\mathrm{diNH_2}) + 0.65\,I(\mathrm{BiBr}) \quad (8)$$

$w_{(\mathrm{mean,class1})} = 1.15 \quad N_1 = 28$

$w_{(\mathrm{mean,class2})} = -1.01 \quad N_2 = 32$

The correct reclassification rate amounts to 91.7% (Class 1,

Table III. Global chemical descriptors appearing in the QSARs

No.	*EHOMO*	*ELUMO*	log*P*
1	−9.0146	−0.6332	2.61
2	−8.3724	−0.3671	2.27
3	−8.2341	0.0005	2.16
4	−8.4704	−0.3989	2.61
5	−9.162	−0.1543	2.79
6	−8.5595	0.0261	3.60
7	−8.5528	0.4059	1.73
8	−8.3483	−0.0286	3.02
9	−8.5626	0.1005	2.95
10	−8.2613	−0.089	1.52
11	−8.7212	−1.0742	0.34
12	−8.4373	0.3141	2.56
13	−8.7767	−0.1195	1.64
14	−8.3253	0.1629	1.91
15	−8.4915	0.4268	2.20
16	−8.3607	0.4333	0.95
17	−8.3096	0.3625	1.01
18	−8.8284	−0.455	0.96
19	−9.3254	−0.8105	0.80
20	−8.1632	0.1491	1.00
21	−8.3083	0.2602	0.76
22	−8.9385	0.0892	2.25
23	−8.5284	−0.4132	1.69
24	−9.4286	−1.5938	0.92
25	−9.0498	−1.0257	0.43
26	−8.0719	0.411	0.48
27	−8.046	0.119	2.38
28	−8.6991	−0.735	1.86
29	−8.0712	0.4025	0.95
30	−8.717	−0.1664	0.20
31	−8.0133	−0.2496	2.39
32	−9.039	−0.7543	0.93
33	−8.6088	0.4153	1.26
34	−8.5707	0.2834	1.01
35	−8.5906	0.1051	1.78
36	−8.3803	0.1761	1.00
37	−8.3819	0.111	1.00
38	−8.5409	0.1441	2.25
39	−8.5193	0.2607	1.48
40	−8.5439	0.2527	1.48
41	−9.0261	−0.3937	1.31
42	−8.1651	0.3817	0.23
43	−8.29	0.4354	0.95
44	−9.3181	−0.5235	2.00
45	−8.2965	0.3429	3.71
46	−8.6024	−0.2779	2.85
47	−7.9895	−0.3544	1.48
48	−9.5438	−1.1895	0.94
49	−9.1996	−1.1264	0.96
50	−8.5709	−0.9914	0.43
51	−8.5894	−0.1632	2.25
52	−10.2645	−1.997	4.25
53	−8.361	0.3996	2.67
54	−8.4687	−0.0638	2.95
55	−9.1425	−1.0677	0.93
56	−8.2927	−0.3932	4.16
57	−8.0935	0.3428	0.20
58	−9.4324	−1.0123	1.22
59	−8.6681	−0.2232	2.58
60	−8.5644	−0.3045	3.09
61	−8.9676	−0.6448	2.72
62	−9.1706	−0.8891	3.73
63	−8.7837	−0.0162	3.21
64	−8.341	−0.3259	3.20
65	−9.1575	−0.5674	0.57
66	−8.6972	0.1085	0.99
67	−9.7645	−0.4175	1.83
68	−9.2655	−0.183	1.69
69	−8.3321	0.4	0.48
70	−8.1015	−0.009	0.60
71	−9.0632	−0.7425	1.04
72	−8.3657	0.4378	2.67
73	−8.4447	−0.4541	1.84
74	−9.2267	−0.6697	0.20
75	−8.4698	0.393	1.73
76	−9.0747	−1.026	0.93
77	−8.7178	−0.3013	2.82
78	−8.5625	−0.4141	1.73
79	−8.2224	−0.1597	1.66
80	−8.1338	−0.0017	3.10
81	−8.9882	−0.6344	2.00
82	−8.3558	0.3482	3.50

EHOMO, energy of the highest occupied molecular orbital; *ELUMO*, energy of the lowest empty molecular orbital; log*P* (P, partition coefficient in the system *n*-octanol/water: values computed from TSAR).

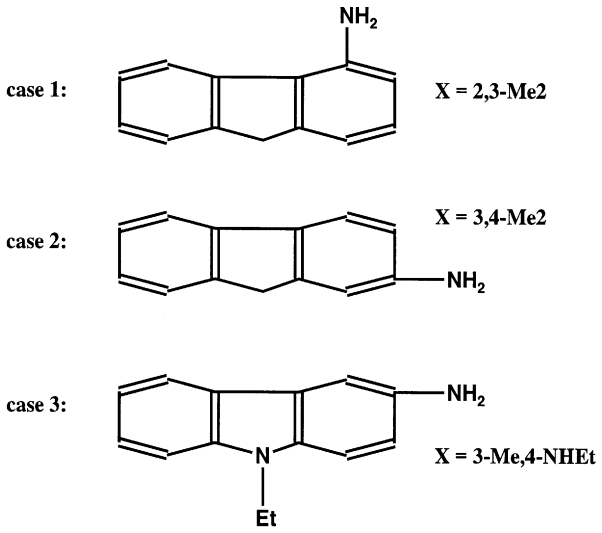

Fig. 3. Treatment of fluorenes.

92.9%; Class 2, 90.6%; misclassified compounds: nos 13, 27, 36, 42 and 75) with a good result for cross validation (all compounds: 83.3%; Class 1, 82.1%; Class 2, 84.4%). If the omitted compound nos 32 and 78 are included in the analysis, the resulting discriminant function does not change but the correct reclassification rate drops to 87.1%.

L(*R*) in Equation 8 can be replaced by *MR*(*R*) without loss of separating power; the resulting discriminant function, Equation 9, now also describes NR2 compounds (without no. 75):

$$w = 0.43\ MR(R) + 0.97\ EHOMO - 1.40\ ELUMO + 0.68\ MR_2 + 0.98\ I(\mathrm{An}) - 0.52\ I(\text{o-}NH_2) - 0.40\ MR_5 + 0.70\ I(\mathrm{diNH_2}) - 0.96\ \log P^{*}I(\mathrm{diNH_2}) + 0.58\ I(\mathrm{BiBr}) \quad (9)$$

$w_{(\mathrm{mean,class1})} = 1.19 \quad N_1 = 30$

$w_{(\mathrm{mean,class2})} = -1.02 \quad N_2 = 35$

The discriminant function in Equation 9 reclassifies 90.8% of the compounds correctly (Class 1, 90.0%; Class 2, 91.4%; misclassified compounds: nos 5, 13, 22, 27, 36 and 42) with an acceptable result in cross validation (all compounds: 83.1%; Class 1, 73.3%; Class 2, 91.4%).

The results obtained for male and female rats resemble each

Table IV. Steric descriptors appearing in the QSARS

No.	MR_5	MR_3	MR_2	MR_6	B_5 (*R*)	*L*(*R*)	*MR*(*R*)
1	0.1	0.56	0.56	0.1	3.13	4.06	1.12
2	0.1	0.8	0.1	0.1	1	2.06	0.1
3	0.1	0.1	0.1	0.1	1	2.06	0.1
4	0.1	0.56	0	0.1	3.13	4.06	1.12
5	0.1	0.6	0.1	0.1	3.43	5.97	2.12
6	0.1	0.1	0.6	0.1	1	2.06	0.1
7	0.1	0.1	0.6	0.1	1	2.06	0.1
8	0.1	0.1	0.1	0.1			1.12
9	0.1	0.1	2.54	0.1	1	2.06	0.1
10	0.1	0.1	0.6	0.6	1	2.06	0.1
11	0.1	0.1	0.74	0.1	2.04	2.87	0.56
12	0.1	0.1	0.1	0.1	1	2.06	0.1
13	0.1	0.1	0.1	0.1	4.04	4.77	1.96
14	0.1	0.1	0.1	0.1	1	2.06	0.1
15	0.1	0.1	0.56	0.56	1	2.06	0.1
16	0.1	0.54	0.56	0.1	1	2.06	0.1
17	0.1	0.1	0.1	0.1	1	2.06	0.1
18	0.1	0.1	0.69	0.1	1	2.06	0.1
19	0.1	0.1	0.1	0.1	3.13	4.06	1.12
20	0.1	0.1	0.6	0.1	1	2.06	0.1
21	0.1	0.1	0.79	0.1	1	2.06	0.1
22	0.1	0.1	0.56	0.56			1.12
23	0.1	0.8	0.8	0.1			1.5
24	0.74	0.1	0.69	0.1	1	2.06	0.1
25	0.1	0.1	0.54	0.1	1	2.06	0.1
26	0.1	0.1	0.1	0.1	1	2.06	0.1
27	0.1	0.1	0.1	0.1	1	2.06	0.1
28	0.1	0.1	0.1	0.1	3.18	4.1	1.56
29	0.1	0.1	0.56	0.1	1	2.06	0.1
30	1.49	0.1	1.25	0.1	1	2.06	0.1
31	0.1	0.56	0.1	0.1	1	2.06	0.1
32	0.1	0.74	0.1	0.1	1	2.06	0.1
33	0.1	0.1	0.1	0.1	1	2.06	0.1
34	0.1	0.1	0.79	0.1	1	2.06	0.1
35	0.1	0.1	0.1	0.1	1	2.06	0.1
36	0.54	0.1	0.6	0.1	1	2.06	0.1
37	0.1	0.1	0.54	0.1	1	2.06	0.1
38	0.1	0.1	0.6	0.1	1	2.06	0.1
39	0.1	0.1	0.56	0.1	1	2.06	0.1
40	0.56	0.1	0.79	0.1	1	2.06	0.1
41	0.1	0.1	0.1	0.1	1	2.06	0.1
42	0.54	0.1	0.79	0.1	1	2.06	0.1
43	0.54	0.1	0.56	0.1	1	2.06	0.1
44	0.1	0.5	0.1	0.1	4.04	4.77	1.96
45	0.1	0.1	0.1	0.1			1.12
46	0.1	0.1	0.1	0.1			1.12
47	0.1	0.8	0.8	0.1	1	2.06	0.1
48	0.1	0.74	0.1	0.1	3.13	4.06	1.12
49	0.74	0.1	0.79	0.1	1	2.06	0.1
50	0.1	0.1	0.74	0.1	1	2.06	0.1
51	0.1	0.1	0.1	0.1	1	2.06	0.1
52	0.1	0.1	0.74	0.74			3
53	0.56	0.1	0.56	0.1	1	2.06	0.1
54	0.1	0.1	0.1	0.1	1	2.06	0.1
55	0.1	0.1	0.28	0.1	1	2.06	0.1
56	0.1	0.8	0.1	0.1	3.11	6.28	2.54
57	0.54	0.1	0.28	0.1	1	2.06	0.1
58	0.1	0.1	0.1	0.1	1	2.06	0.1
59	0.1	0.1	0.1	0.1	3.13	4.06	1.12
60	0.1	0.1	0.1	0.1	1	2.06	0.1
61	0.1	0.1	0.1	0.1	3.13	4.06	1.12
62	0.1	0.56	0.1	0.1	3.67	4.65	1.12
63	0.1	0.1	0.1	0.1	1	2.06	0.1
64	0.1	0.1	0.6	0.1	1	2.06	0.1
65	0.1	0.1	0.1	0.1	3.13	4.06	1.12
66	0.1	0.1	0.1	0.1	3.13	4.06	1.12
67							
68							
69	0.1	0.1	0.54	0.1	1	2.06	0.1
70	0.1	0.1	0.54	0.1	1	2.06	0.1
71	0.09	0.54	0.09	0.09	1	2.06	0.1
72	0.1	0.1	0.56	0.56	1	2.06	0.1
73	0.1	0.1	0.1	0.1			1.12
74	0.1	0.74	0.1	0.1	1	2.06	0.1
75	0.1	0.1	0.1	0.1	1	2.06	0.1
76	0.74	0.1	0.28	0.1	1	2.06	0.1
77	0.1	0.1	0.6	0.6	1	2.06	0.1
78	0.1	0.56	0.1	0.1	1	2.06	0.1
79	0.1	0.1	0.79	0.1	1	2.06	0.1
80	0.1	0.1	0.56	0.1	1	2.06	0.1
81	0.6	0.69	0.6	0.1	1	2.06	0.1
82	0.1	0.1	0.56	0.1	1	2.06	0.1

MR_i, molar refractivity of a substituent in position *i* of the aniline core (for conventions, see text). $B_5(R)$ and $L(R)$, Sterimol parameter for *R* in NHR; $MR(R)$, molar reactivity of *R* in NHR and of R1 and R2 in NR1R2.

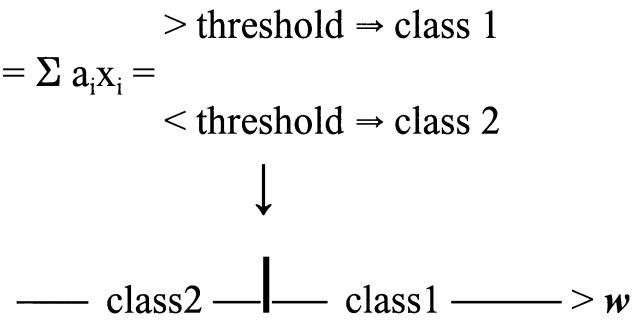

Fig. 4. Discriminant function in a two-classes case.

other and are similar to the overall carcinogenicity results. Of key importance for class separation are electronic properties as expressed by *EHOMO* and *ELUMO*, the type of ring system, and substitution in the ortho-position as well as at the amino nitrogen. The probability of a compound being assigned to the active class increases with increasing values of *ELUMO*, decreasing values of *EHOMO*, decreasing bulk of substituents in position 2 (ortho-position), decreasing length (or bulk) of substituents at the amino nitrogen and increasing number of aromatic rings (anilines have a distinctively lower probability of being active than biphenyls, fluorenes or naphthalenes). An important feature promoting carcinogenic potency also is the occurrence of an amino group in the ortho-position to the functional amino group. Of lesser importance are the variables $I(\mathrm{diNH_2})$, I(BiBr), MR_5, and the cross product $\log P{*}I(\mathrm{diNH_2})$. The $I(\mathrm{diNH_2})$ term seems to indicate that compounds with more than one amino group are intrinsically a little less active than compounds with only one amino group. This effect is counterbalanced by increasing hydrophobicity. Since the $\log P{*}I(\mathrm{diNH_2})$ term outperforms the $I(\mathrm{diNH_2})$ term the general message is that compounds with more than one amino group have an increased probability of being carcinogenic with increasing hydrophobicity. Other things being equal, the occurrence of a bridge between the two aromatic rings in biphenyls seems to decrease and bulk in meta position seems to increase carcinogenicity potency.

It becomes obvious again that the key factors differentiating between active and inactive compounds on the one hand, and those governing potency within the group of active compounds are different. The most pronounced differences are with respect

to the importance of hydrophobicity and the directionality of electronic effects (see Discussion).

Mice carcinogenicity

In contrast to the overall results and to the rat carcinogenicity, only a very moderate class separation (reclassification rate ≈ 74%) with no stability in cross validation was achieved for mice carcinogenicity if all compounds were included. Therefore compounds had to be omitted not only in order to obtain an improved result, but also since no meaningful result at all would otherwise be obtained. In fact, the experience accumulated from years of study in the QSAR field has indicated that outliers are most often compounds for which the experimental data are unreliable, or that do not follow the same mechanism of action as other chemicals in the set (12,13). The omitted compounds were no. 34 for female and male mice plus nos 50 and 66 in the case of female mice and nos 42 and 75 in the case of male mice. For female carcinogenicity, the description of N-substitution by $L(R)$ and $B_5(R)$ was not essential, so that only the discriminant function with the simple indicator I(NR) is presented. For male mice, on the contrary, replacement of $L(R)$ and $B_5(R)$ by simpler descriptors brings about a sharp loss of separating power; therefore, only the discriminant function with these variables will be presented (only four N-disubstituted compounds were lost this way).

For female mice carcinogenicity, the following discriminant function reclassifies 85.7% of the compounds correctly (Class 1, 87.9%; Class 2, 83.3%; misclassified compounds: nos 7, 30, 38, 42, 49, 69, 70, 75 and 77) and is of acceptable stability in cross validation (all compounds: 81.0%; Class 1, 84.8%; Class 2, 76.7%):

$$w = -0.47\ I(NR) + 1.38\ \log P^{*} I(\mathrm{monoNH_2}) + 1.68\ \log P^{*} I(\mathrm{diNH_2}) - 0.37\ I(\mathrm{An})(4.3.1) + 0.33\ I(\text{o-NH2}) - 0.55\ MR_5 - 0.45\ I(\mathrm{BiBr}) \quad (10)$$

$$w_{(\mathrm{mean,class1})} = -0.92 \quad N_1 = 33$$
$$w_{(\mathrm{mean,class2})} = 1.01 \quad N_2 = 30$$

For male mice the discriminant function in Equation 11 is obtained:

$$w = -1.96\ L(R) + 1.69\ B_5(\mathrm{R}) - 0.83\ EHOMO + 0.97\ ELUMO - 1.22\ I(\mathrm{An}) + 0.73\ I(\text{o-NH2}) + 0.59\ MR_3 + 0.69\ MR_5 + 0.77\ MR_6 - 0.76\ I(\mathrm{diNH_2}) + 1.09\ \log P^{*} I(\mathrm{diNH_2}) - 0.79\ I(\mathrm{BiBr}) \quad (11)$$

$$w_{(\mathrm{mean,class1})} = -1.11 \quad N_1 = 25$$
$$w_{(\mathrm{mean,class2})} = 1.16 \quad N_2 = 24$$

The discriminant function in Equation 11 shows a good reclassification rate (all compounds: 89.8%; Class 1, 96.0%; Class 2, 83.3%; misclassified compounds: nos 7, 16, 49, 66 and 71) and stability in cross validation (all compounds: 83.7%; Class 1, 96.0%; Class 2, 70.8%).

The results were similar to the overall results and to the rat carcinogenicity results, with the difference that electronic effects, as expressed by *EHOMO* and *ELUMO*, were of smaller importance. In the case of male mice carcinogenicity, a correct reclassification rate of 79.6% was still obtained if *EHOMO* and *ELUMO* were eliminated from discriminant function Equation 11, and for female mice carcinogenicity no electronic terms were observed. There is, however, a significant multiple correlation between the variables appearing in Equation 10 and *EHOMO* ($r = 0.64$) as well as (*ELUMO* – *EHOMO*) ($r = 0.52$) so that some electronic effect may well be hidden behind these variables. The direction of the electronic effect for male mice carcinogenicity is the same as observed for the other carcinogenicity scales (the probability of a compound being assigned to the active class increases with decreasing values of *EHOMO* and increasing values of *ELUMO*) and, thus, again opposite to the effect observed for the gradation of potency in mice (Equation 2). Very important for class separation is the type of ring system: I(An) alone already reclassifies >80% of the Class 1 compounds (but only ~50% of the Class 2 compounds) correctly so that anilines have a distinctly lower probability of being active than biphenyls, fluorenes or naphthalenes. Hydrophobicity and the number of amino groups also influence carcinogenicity, but this influence is only of secondary importance. For male mice, the probability of a compound being assigned to the active class decreases if more than one amino group is present; this effect is counterbalanced by increasing hydrophobicity of the poly-amino compounds. An increase with hydrophobicity of both, compounds with only one and with more than one amino group, is evident in the case of female mice; this effect is more pronounced in the latter group of compounds. The difference between the discriminant functions in Equations 10 [the terms $\log P^{*} I(\mathrm{monoNH_2})$ and $\log P^{*} I(\mathrm{diNH_2})$ occur but not the term $I(\mathrm{diNH_2})$] and 11 [the terms $\log P^{*} I(\mathrm{diNH_2})$ and $I(\mathrm{diNH_2})$ occur but not $\log P^{*} I(\mathrm{monoNH_2})$] are probably due to fluctuations in the data structure, as these variables are highly correlated for both the data set for female as well as for male mice.

For female mice, substitution at the amino nitrogen is unfavorable for (decreases) carcinogenic potency as was also seen for female and male rats. For male mice, however, the situation is somewhat more complex and corresponds to the picture obtained for overall carcinogenicity. With regard to mono-substitution at the amino nitrogen, long substituents are unfavorable while substituent width seems to be allowed. The $B_5(R)$ term in discriminant function Equation 11 can be replaced by a $L(R)^2$ term without changing anything else, indicating a length optimum for the N-substituent at about $L(R)_0 \approx 3.4$. As it is, of course, not possible to increase $B_5(R)$ substantially without also increasing $L(R)$, the combination of $L(R)$ and $B_5(R)$ as well as of $L(R)$ and $L(R)^2$ do, in fact, tell the same story. The length optimum is close to the length of COMe ($L = 4.06$) (see Discussion).

An important feature also promoting carcinogenic potency is the occurrence of an amino group in the ortho-position to the functional amino group. Substitution (bulk) in the meta-position supports carcinogenic activity as was also seen for the other carcinogenicity scales; again, this effect is opposite to that reflected by Equations 1 and 2. In the ortho-position a similar effect is seen for male mice carcinogenicity as already observed for the overall carcinogenicity. The MR_6 term codes for 2,6-Me_2 or 2,6-Cl_2 substitution, and this pattern seems to be advantageous for carcinogenic potency in keeping with the evidence presented in Equation 7. Finally, a bridge between the phenyl rings in the biphenyls seems to decrease carcinogenicity. The important point again is that the effects governing yes/no activity and the gradation of potency within the active compounds are different (see Discussion).

Discussion

Comparing and interpreting the QSAR models

Discriminant analysis leads to discriminant functions with high separating power and good stability in cross validation for all

carcinogenicity scales. Similar properties affect yes/no activity and the gradation of carcinogenic potency within the group of active carcinogens. However, the relative importance and even the direction of some of these effects are very different, which explains why the QSARs in Equations 1 and 2 cannot differentiate between carcinogens and non-carcinogens. The first difference concerns hydrophobicity, which is a key factor for the gradation of carcinogenic potency, but only of small importance for yes/no activity. The reverse is true for electronic properties (*EHOMO*, *ELUMO*) which show only a small effect for the gradation of potency, but a pronounced effect (especially *ELUMO*) for yes/no activity (especially for overall and rat carcinogenicity). In addition, the directionality of the electronic effect is different. While, according to Equation 1, the degree of carcinogenic potency in mice increases with increasing values of *EHOMO* and decreasing values of *ELUMO*, the reverse is true for the probability of a compound belonging to the class of carcinogens. A possible explanation is that too high a reactivity might draw the molecules into a different route of reactions (metabolic inactivation, or they do not find the right target/interact with a different target) preventing them becoming carcinogenic. Thus, given a set of molecules, which have identical properties with respect to all other terms appearing in the discriminant functions, a reactivity limit may exist above which a compound will become non-carcinogenic. This does not mean, of course, that all molecules exceeding this limit are non-carcinogenic, as yes/no activity also depends on additional properties. In other words, for a hypothetical set of compounds which have identical properties except for *EHOMO* and *ELUMO*, there is some limiting value of the combination of these quantities above which the molecules tend to become inactive. Below this value the degree of carcinogenicity increases with *EHOMO* and decreases with *ELUMO*.

Another difference exists with respect to substitution at the amino nitrogen. The effect for the gradation of potency is only moderate, while a strong effect exists for yes/no activity. The general tendency is that bulk at the nitrogen blocks carcinogenic activity. However, for overall and male mice carcinogenicity a length optimum exists which is close to the length of the COMe group. This probably reflects the fact that the first step of metabolic activation involves N-hydroxylation and/or N-acetylation (7).

According to Equation 1, bulk in the ortho-position decreases potency. This is also seen in the discriminant functions for rat carcinogenicity (MR_2 term) but not in those for overall and mice carcinogenicity, where bulk in position 6 increases the probability of a compound falling into the active class (MR_6 term). The results in Equation 1 show that it depends on the nature of an ortho-substituent whether it will decrease or increase carcinogenic potency. According to the conventions used, the appearance of a substituent in position 6 implies 2,6-disubstitution. There are four compounds of this type in the data set with either 2,6-Me_2 or 2,6-Cl_2 (nos 8, 13, 65 and 70) which all belong to the active class with respect to overall and male mice carcinogenicity. The real meaning of the MR_6 term is that this substitution pattern obviously supports carcinogenic potency which is in keeping with the statement 'substitution of a chloro group or methyl group or methoxy group ortho to the amino group often enhances potency' (7). An inhibition is reported only for larger substituents.

Equation 1 (but not Equation 2) also shows a steric effect in meta-position: bulk decreases potency. For yes/no activity, on the contrary, a supporting effect of bulk in this position is observed in the discriminant functions for all carcinogenicity scales. It must be noted that the variation of bulk in the meta position(s) is only limited, so that the meta bulk terms are not very well supported. A possible explanation for the apparently different effects of bulk in meta with respect to yes/no activity and the gradation of mice carcinogenicity may be that bulk in this position affects metabolism and the interaction of the ultimate carcinogen with its target in a different way.

A further difference between the QSARs for the gradation of potency within active compounds and yes/no activity regards the type of aromatic ring system which has a very strong effect for yes/no activity but is of only small importance in the case of the gradation of rat carcinogenicity (Equation 2) and of no importance for the gradation of mice carcinogenicity (Equation 1). For all carcinogenicity scales, the probability of a compound being assigned to the active class is smallest if only one aromatic ring is present (anilines).

As for the gradation of potency, the presence of a bridge between the two phenyl rings in biphenyls is also unfavorable for yes/no activity. A smaller effect influencing yes/no activity, but not the gradation of carcinogenic potency, is the activity supporting role of a NH_2 group adjacent to the functional amino group.

QSAR models in perspective

A remarkable aspect of the present result is that the QSAR models are in keeping with, and can be interpreted based on what is known about, the mechanisms of action of aromatic amines. Aromatic amines have to be metabolized to reactive electrophiles in order to exert their carcinogenic potential. For aromatic amines and amides, this typically involves an initial N-oxidation to *N*-hydroxyarylamine and *N*-hydroxyarylamide (7,9). This is in agreement with the importance of the chemical reactivity parameters (*EHOMO* and *ELUMO*) in the QSAR models. In particular, EHOMO is a parameter for oxidation reactions. Moreover, steric factors (bulk, shape) are critical in the interaction between enzymes and chemicals to be metabolized (13). All the above parameters appear to make the difference between the amines that can be processed by the cellular machinery and those that cannot. On the contrary, hydrophobicity [which is normally a fundamental parameter for transport and ease of interaction with the enzymes (13)] appears to have a primary role for the gradation of the carcinogenic potency (Equations 1 and 2), but not for setting the threshold between carcinogenic and non-carcinogenic amines. Interestingly, the mutagenic properties of the aromatic amines also pointed to a similar picture: the patterns of molecular determinants for the potency and the yes/no activity were different, and were analogous to those found here for their carcinogenicity (2,14). This evidence also represents an indirect proof of the similarity of the mechanisms by which the aromatic amines act in Salmonella typhimurium (mutagenicity) and in rodents (carcinogenicity).

Another point to be remarked is that the QSAR models obtained can be used directly for estimating the carcinogenicity of non-heterocyclic aromatic amines for which experimental carcinogenicity data are not available. With the QSARs in Benigni *et al*. (1) and the present results, a two-step prediction of carcinogenicity of aromatic amines seems to be possible: (i) step 1, yes/no activity from the discriminant functions; and (ii) step 2, if the answer from step 1 is yes then prediction of the degree of potency from the Hansch equations in Benigni

et al. (1). Thus, the QSAR models can contribute to the following: the direct synthesis of safer chemicals; estimation of the risk posed by amines in the environment; setting priorities for further experimentation, thus also reducing the use of experimental animals. Even though the mathematical models provide estimations and cannot replace the experimental results (when necessary), the goodness of fit of the present models point to a remarkable level of reliability for their practical use.

A critical aspect of QSAR modeling is the availability of a sufficient number of good quality data, and the fulfillment of certain requirements, such as sampling the chemicals in such a way that the chemical descriptors are poorly intercorrelated. This to get clearer responses from the analysis (12,13). Unfortunately, these requirements are seldom fulfilled in toxicological QSAR analyses, notably in the case of rodent carcinogenicity results: the bioassays are too expensive and time consuming, and they are planned according to criteria (extent of use of the chemicals, specific scientific interest) different from those typical of QSAR. One has to use the data that are available in the literature. In the present work, we were in a far better situation than most of the other QSAR studies of carcinogenicity, since the class of aromatic amines is, by far, the most extensively bioassayed: the number of chemicals was sufficient for a thorough QSAR analysis, and out of the 79 chemicals actually analysed, the great majority (55) were derived from the same laboratory and were generated with the same protocol by the NTP. At the same time, we are aware that the different experimental origin of the remaining 24 chemicals (retrieved in the Gold's database) may add some 'noise' to the data set. However, the QSAR modelling we made was largely successful. The QSAR models were both good from a statistical point of view, and—most important—were coherent and meaningful from a mechanistic point of view. The cogency of our results was even more supported by the fact that the QSAR models for the rodent carcinogenicity of the aromatic amines were quite similar to those obtained previously by us and other investigators (e.g. Corwin Hansch) for Salmonella mutagenicity (2,14). This is in agreement with the accepted notion that the basic steps of the action mechanism of aromatic amines are similar in the two experimental systems. This means that the QSAR modelling was able to highlight the general trends underlying the action of the aromatic amines, and was not confused by the experimental noise.

The successful modeling of in vivo data provided in this and in our previous paper deserves further comment. Whereas experimental results from *in vitro* systems are generally considered reliable enough for the building models, the quality of in vivo data is often questioned. On the contrary, the robustness and intepretability of our results show that in vivo data can be successfully modelled. The possibility demonstrated in this paper of defining not only the molecular determinants of the gradation of potency, but also the existence of a marked chemical difference between carcinogenic and non-carcinogenic aromatic amines has further implications for the rodent carcinogenicity assay. In opposition to the claims that many positive results in the bioassay are artifacts due to aspecific toxic effects of the high doses employed (15), the chemical difference between active and inactive amines strongly supports the wide range of arguments that toxicity has no, or a minor, role in the bioassay results (16–18). Overall, this evidence supports the reliability of the traditional rodent carcinogenicity assay.

References

1. Benigni,R., Giuliani,A., Franke,R. and Gruska,A. (2000) Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev.*, **100**, 3697–3714.
2. Benigni,R., Passerini,L., Gallo,G., Giorgi,F. and Cotta-Ramusino,M. (1998) QSAR models for discriminating between mutagenic and nonmutagenic aromatic and heteroaromatic amines. *Environ. Mol. Mutagen.*, **32**, 75–83.
3. Vineis,P. and Pirastu,R. (1997) Aromatic amines and cancer. *Cancer Causes Control*, **8**, 346–355.
4. Dolin,P.J. (1992) A descriptive study of occupation and bladder cancer in England and Wales. *Br. J. Cancer*, **65**, 476–478.
5. Rubino,G.F., Scansetti,G., Piolatto,G. and Pira,E. (1982) The carcinogenic effect of aromatic amines, an epidemiologic study on the role of *o*-toluidine and 4,4'-methylene bis (2-methylaniline) in inducing bladder cancer in man. *Environ. Res.*, **27**, 241–254.
6. Stasik,M.J. (1988) Carcinomas of the urinary bladder in a 4-chloro-*o*-toluidine cohort. *Int. Arch. Occup. Environ. Health*, **60**, 21–24.
7. Lai,D.Y., Woo,Y.T., Argus,M.F. and Arcos,J.C. (1996) Cancer risk reduction through mechanism-based molecular design of chemicals. In DeVito,S.C. and Garrett,R.L. (eds) *Designing Safer Chemicals. Green Chemistry for Pollution Prevention*. American Chemical Society, Washington, DC, pp. 1–13.
8. Kadlubar,F.F. (1994) DNA adducts of carcinogenic aromatic amines. In Hemminki,K., Dipple,A., Shuker,D.E.G., Kadlubar,F.F., Segerback,D. and Bartsch,H. (eds) *DNA Adducts, Identification and Biological Significance*. IARC, Lyon, pp. 199–216.
9. Beland,F.A. and Kadlubar,F.F. (1990) Metabolic activation and DNA adducts of aromatic amines and nitroaromatic hydrocarbons. In Cooper,I.C.S. and Grover,P.L. (eds) *Chemical Carcinogenesis and Mutagenesis*. Springer Verlag, Berlin, pp. 265–325.
10. Gold,L.S., Manley,N.B., Slone,T.H. and Rohrbach,L. (1999) Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environ. Health Perspect.*, **107** (suppl. 4), 527–600.
11. Selkirk,J.K. and Soward,S.M. (1993) Compendium of abstracts from long-term cancer studies reported by the National Toxicology Program from 1976 to 1992. *Environ. Health Perspect*., **101**, 1–249.
12. Franke,R. (1984) *Theoretical Drug Design Methods*. Elsevier, Amsterdam.
13. Hansch,C. and Leo,A. (1995) *Exploring QSAR. 1. Fundamentals and Applications in Chemistry and Biology.* American Chemical Society, Washington, DC, vol. 1.
14. Debnath,A.K., Debnath,G., Shusterman,A.J. and Hansch,C. (1992) A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100. *Environ. Mol. Mutagen.*, **19**, 37–52.
15. Ames,B.N. and Gold,L.S. (1990) Chemical carcinogenesis: Too many rodent carcinogens. *Proc. Natl Acad. Sci. USA*, **87**, 7772–7776.
16. Huff,J. (1999) Value, validity and historical development of carcinogenesis studies for predicting and confirming carcinogenic risks to humans. In Kitchin,K.T. (ed.) *Carcinogenicity: Testing, Predicting and Interpreting Chemical Effects*. Marcel Dekker, New York, pp. 21–123.
17. Huff,J. (1993) Absence of morphologic correlation between chemical toxicity and chemical carcinogenesis. *Environ. Health Perspect.*, **101**, 45–33.
18. Tomatis,L., Huff,J., Hertz-Picciotto,I., Sandler,D.P., Bucher,J., Boffetta,P., Axelson,O., Blair,A., Taylor,J., Stayner,L. and Barrett,J.C. (1997) Avoided and avoidable risks of cancer. *Carcinogenesis*, **18**, 97–105.

Received June 30, 2000; revised December 29, 2000; accepted May 22, 2001