Bull. Environ. Contam. Toxicol. (1984) 32: 354–362
© 1984 Springer-Verlag New York Inc.

Bulletin of Environmental Contamination and Toxicology

Molecular Connectivity of Phenols and Their Toxicity to Fish

Lowell H. Hall[1] and Lemont B. Kier[2]

[1]Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170, and
[2]Department of Pharmaceutical Chemistry, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA 23298

Reliable studies on the toxicity of a large number of compounds have recently become available to the scientific community from the Environmental Research Laboratory at Duluth, Minnesota, The Stanford Research Institute and the University of Wisconsin at Superior. This extensive database makes it possible to contact QSAR studies for the purpose of establishing structural relationships with toxicity in order to make predictions and possibly shed light on the mechanisms involved. In our laboratories we have utilized this data in several studies on the QSAR of toxicity of compounds of commercial interest which may influence the environment (Hall and Kier 1983).

One group of compounds, the phenols, are of particular interest due to their wide use and potential for environmental contamination. The purpose of this study is to relate the relative toxicity to the structure in an effort to make predictions about the toxic potentiality of other phenols.

METHODS AND MATERIALS

Flow-through acute toxicity studies were conducted using laboratory reared fathead minnows Pimephales promelas as described earlier (Kier and Hall in press). All experiments were conducted at 25°C using minnows which were 30 to 35 days old. Water from Lake Superior was carefully prepared and monitored for all studies. Each test used 50 fish and the number of dead were determined on a regular schedule. The lethal concentration for a 50% kill of the sample was determined statistically as LC50. The negative logarithm pLC50 is used in the QSAR analysis and the computed values are shown in Table 1 along with the compound names. The toxicity values range over three orders of magnitude with the pLC50 values running from 3.21 to 6.20. There are 15 different substituents ranging in size from chloro, bromo, methyl to phenyl and nonyl.

*Correspondence and reprint requests

In this study the biological activity, pLC50, expressed as the logarithm of the measured lethal concentration, is related to the molecular structure of the compounds. For the quantitation of the structure we have used molecular connectivity indexes (Hall and Kier 1976). The advantage of this approach over the use of a physical property is the potential for the direct interpretation of the structure in terms of familiar molecular fragments. The molecular connectivity indexes encode information about molecular size, skeletal branching, unsaturation, heteroatom content and intramolecular distances as in the case of benzene ring substituents.

The molecular connectivity indexes arise from the assignment of delta values to atoms, other than hydrogen, in a molecule. The simple delta value, δ, is a count of the number of sigma bonds (other than to hydrogen) emanating from that atom and is equal to the number of skeletal bonds: $\delta = \sigma - h$, where σ is the number of sigma bonds and h is the number of hydrogen atoms bonded to the atom.

The molecule is then dissected into substructures of one, two or more contiguous bonds (paths). Each path is defined by the delta values describing each atom in the substructure. A term, S_k, for each path (substructure of one bond) is calculated according to the algorithm.

$$S_k = (\delta_i \delta_j)_k^{-1/2}$$

where δ_i and δ_j are delta values assigned to the two atoms of bond k (or path of the first order). The molecular connectivity index of the first order, $^1\chi$, is computed as the sum of all the first order path terms in the molecular skeleton:

$$^1\chi = \Sigma_k S_k \quad \text{for all bonds in the skeleton}$$

This simple first order index has been shown to rank molecules according to their size and branching and to be closely related to many of their physical properties (Hall and Kier 1978 and 1981). Higher order indexes, based on larger substructures, encode more complex aspects of structure and often appear in multivariate regression analyses of molecules for physical and biological properties (Kier 1980).

A second class of molecular connectivity indexes arises from the assignment of atom delta values based upon a count of all atom valence electrons other than those involved in covalent bonds to hydrogen. The valence delta value δ^v is defined for first row elements as follows:

$$\delta^v = \sigma + p + n - h = Z^v - h$$

where p is the number of electrons in pi bonds, n is the number electrons involved in lone pairs and Z^v is the number of valence electrons. For higher row elements core electrons (counted as $Z - Z^v - 1$) are explicitly included:

$$\delta^v = \frac{Z^v - h}{Z - Z^v - 1}$$

The valence delta values, δ^v, are employed just like the simple delta values to compute a family of valence molecular connectivity indexes, ${}^m\chi_t^v$, of various orders m the number of contiguous bonds in the corresponding substructural fragment, and type t of substructure such as path(P), cluster(C), path-cluster(PC). Thus, simple chi indexes ${}^m\chi_t$ (based on simple deltas δ) reflect skeletal arrangement only whereas valence chi indexes ${}^m\chi_t^v$ (based on valence deltas δ^v) also encode atom type and contain electronic information (Kier and Hall, 1981) about the molecule.

RESULTS AND DISCUSSION

The toxicity of the 25 phenols was analyzed using standard regression methods against the computed molecular connectivity indexes. The best single variable equation was found to be as follows:

$$pLC50 = 1.079\ {}^3\chi_P^v + 2.528 \qquad (1)$$

$$r = 0.903,\ s = 0.347, F = 101,\ n = 25$$

The best equation using two variables is as follows:

$$pLC50 = 0.205\ {}^1\chi + 0.906\ {}^3\chi_P^v + 1.786 \qquad (2)$$

$$r = 0.934,\ s = 0.296,\ F = 75,\ n = 25$$

The connectivity variables examined in the regression analysis include the following: ${}^0\chi$, ${}^1\chi$, ${}^2\chi$ and the path indexes ${}^3\chi_P$ through ${}^6\chi_P$, ${}^3\chi_C$, ${}^4\chi_{PC}$, and ${}^6\chi_{CH}$ along with the corresponding valence indexes for a total of 20. The correlation of the two variables ${}^1\chi$ and ${}^3\chi_P^v$ show that they are very nearly independent, $r^2 = 0.26$.

The use of a third variable may not be justified on the simple statistical grounds that the ratio of observations to variables would be reduced below 10. Further, an analysis of the residuals from eq. 2 (see Table 1) shows a random distribution with no trend to nonlinearity. An estimate of the precision of the data is about ± 0.15 log units (Kier and Hall in press); hence, equation 2 produces a standard deviation about twice that in the experimental data. From Table 1 it is evident that eq. 2 predicts no value which differs from the observed value by more than twice the regression standard deviation.

Table 1. Molecular Connectivity Indexes and Toxicities for Substituted Phenols

No.	Compound Name	$^1\chi$	$^3\chi_p^v$	-log LC_{50} Obs[a]	-log LC_{50} Calc[b]	Res
1	Phenol	3.394	0.756	3.51	3.17	0.34
2	2-Chloro-	3.805	1.172	4.02	3.63	0.39
3	2-Allyl-	4.843	1.356	3.93	4.01	-0.08
4	4-Nitro-	4.698	1.061	3.36	3.71	-0.35
5	3-Methoxy-	4.326	1.053	3.21	3.63	-0.42
6	4-Butyl-	5.326	1.755	4.47	4.47	0.00
7	4-Tertbutyl-	4.999	1.733	4.46	4.38	0.08
8	4-Pentyl-	5.826	2.005	5.18	4.80	0.38
9	4-Tertpentyl-	5.560	2.533	4.82	5.22	-0.40
10	2-Phenyl-	6.377	1.989	4.45	4.90	-0.45
11	4-Phenylazo-	7.343	1.887	5.26	5.00	0.26
12	4-Nonyl-	7.826	3.005	6.20	6.11	0.09
13	2,4-Dichloro-	4.198	1.448	4.30	3.96	0.34
14	4-Chloro-3-Methyl-	4.198	1.566	4.27	4.07	0.20
15	2,4-Dimethyl-	4.198	1.352	3.86	3.87	-0.01
16	2,6-Dimethyl-	4.215	1.439	3.74	3.95	-0.21
17	3,4-Dimethyl-	4.198	1.490	3.94	4.00	-0.06
18	4-Amino-2-Nitro-	5.109	1.196	3.65	3.92	-0.27
19	2,4-Dinitro-	6.020	1.367	4.04	4.26	-0.22
20	2,4,6-Trichloro-	4.609	1.777	4.33	4.34	-0.01
21	2,4,6-Tribromo-	4.609	2.650	4.70	5.13	-0.43
22	2-Methyl-4,6-Dinitro-	6.430	1.673	4.99	4.62	0.37
23	2,3,4,5-Tetrachloro-	5.037	2.826	5.72	5.38	0.34
24	Pentachloro-	5.464	3.446	6.06	6.03	0.03
25	1-Naphthol	5.377	1.767	4.53	4.49	0.04

[a] Data taken from ref 1. [b] Calculated from eq. 2.

When a regression equation based on one or two variables is developed from a larger set of variables, there is always a question of statistical significance (Topliss and Costello 1972, Kier and Hall 1977). Thus, as a further check of the quality of these regressions, two statistical studies were performed.

In the first study the chi variables were replaced by random numbers. The same programs used to generate eqs. 1 and 2 were then employed to obtain the best correlation with the random numbers as variables. This procedure was then repeated for a total of 100 times. For best correlation based on one variable the largest correlation coefficient obtained with the random numbers is $r = 0.596$. Equation 1 shows that the use of chi variables is significantly different ($r = 0.903$) than randomly chosen variables. With the use of two variables the random number variables yielded only two correlation coefficients greater than 0.70, the largest being $r = 0.717$. The total number of two-variable correlations actually examined in this study is 19,000. We conclude that the probability of random correlations achieving an $r > 0.70$ is two in 19000 or about 0.011%. The probability for $r > 0.90$ is even smaller. Equation 2 is seen, then, to be very significant statistically.

In the second study the statistical stability of the final regression equation (eq 2) is examined. The details of any regression equation depend upon the actual set of observations used in the regression analysis. The effect of varying the data set is simulated by leaving out observations and recomputing the regression equation. This procedure may be repeated by randomly selecting observations to be deleted until each observation has been deleted at least once. The stability of the regression model may be judged by the degree of variance in the statistical information computed for all cases in which observations have been deleted.

The results of the statistical stability study are summarized in Table 2. In the left hand column are results from the standard regression analysis which yielded eq. 2. The right hand columns are the result of deleting 3 observations (12% of the data), selected on a random basis, 30 times. The result reported in each case is the average and standard deviation for those 30 runs. The very favorable comparison of the right hand side with the left hand side indicates that the regression model is very stable with respect to choice of data set. This is especially indicated by the low dispersion (standard deviation of the average) for each of the statistical categories listed in Table 2. (Similar results were also obtained when a larger percentage of the data,up to 25%, were deleted.) For this present study each observation

Table 2. Information for Test of Statistical Stability of Regression Model

	Original Model (no deletions)		3 Deletions per Run (30 runs)	
	Regression	Standard Deviation	Average	Standard Deviation
Correlation Coefficient	0.934		0.933	0.010
Standard Deviation, s	0.296		0.297	0.011
Coefficient of $^1\chi$, a_1	0.205	0.004	0.204	0.026
Coefficient of $^3\chi p$, a_2	0.906	0.011	0.906	0.038
Constant, a_0	1.786	0.295	1.787	0.107
Average residual	0.231		0.254	0.171
Residuals less than one standard deviation	56.0%		51.1%	
Residuals between one and two standard deviations	44.0%		48.9%	
Residuals greater than two standard deviations	0.0%		0.0%	

was deleted at least once; no observation was deleted more than 4 times and the typical number of times an observation was deleted in the 30 runs is 3 or 4.

Further, for each observation deleted (that is, not included in the current regression calculation) a value was computed for the activity. This "predicted" value was compared to the observed and the residual (obs-calc) stored. The analysis of these residuals is reported at the bottom of Table 2. Once again the comparison to the original regression model (eq. 2) is most favorable.

These results taken all together support the idea that eq. 2 is statistically significant and may be used as a basis of estimation of the toxicity of phenols.

Equation 2 relates the toxicity to two connectivity variables, $^1\chi$ and $^3\chi p$. This connectivity equation lends itself to a direct interpretation of the important structural characteristics which influence toxicity in this data set. The $^1\chi$ index with a positive coefficient reflects the fact that an increase in molecular size results in an increase in toxicity. The $^1\chi$ index in this equation is the simple index rather than the valence index, indicating that only the

number of atoms and their skeletal connections are encoded here. Thus, this index contains information that the number of skeletal atoms and their degree of branching influences toxicity. The nature of heteroatoms and their impact on toxicity is not contributed by this index for this data set. An examination of the contribution of $^{1}\chi$ to toxicity in eq. 2 reveals that the simple index contributes about 40% to the variation in toxicity whereas the $^{3}\chi_{P}^{v}$ index contributes the remaining 60%. Thus, the influence of molecular size is significant but not dominant.

It is possible to examine subsets of the data using only the first order index. There are 12 compounds with substituents consisting only of hydrocarbon groups, along with the parent molecule phenol. The first order valence index $^{1}\chi^{v}$ gives an excellent correlation for these 12 observations:

$$pLC50 = 0.631\ ^{1}\chi^{v} + 2.03$$

$$r = 0.977,\ s = 0.17,\ F = 208,\ n = 12$$

These compounds include allylphenol, phenylphenol and naphthol in addition to phenol and those compounds with alkyl substituents. Correlation with $^{1}\chi^{v}$ is significantly better than with the number of atoms ($s = 0.27$). When the subset of compounds is narrowed to include only the alkyl side-chain phenols, both the simple and valence indexes yield the same statistics since they contain essentially the same structural information. For saturated hydrocarbons, as in the alkyl side-chains, $^{1}\chi = {}^{1}\chi^{v}$.

$$pLC50 = 0.643\ ^{1}\chi + 1.21$$

$$r = 0.989,\ s = 0.14,\ F = 298,\ n = 9$$

The $^{3}\chi_{P}^{v}$ index in eq. 2 has a richer information content than does $^{1}\chi$. Further, this information is more important to toxicity than is the $^{1}\chi$ index as shown by the moderately good correlation in eq. 1 as compared to the correlation of just $^{1}\chi$ with toxicity, 0.61. The $^{3}\chi_{P}^{v}$ index encodes two basic elements of structural information (Hall and Kier in press). The first is the indication that valence definition of the delta values is influential in describing toxicity. The presence of heteroatoms in the first row of the periodic chart lowers toxicity relative to the methyl group whereas substituents in higher quantum levels produce a greater toxicity. Compare 2,4-dichlorophenol (cpd. 13, 4.30) with 2,4-dimethylphenol (cpd. 15, 3.86) and 2,4,6-tribromophenol (cpd. 21, 4.70) with 2,4,6-tricholorophenol (cpd. 20, 4.33).

The $^{3}\chi_{P}^{v}$ index is larger when ring substituents are adjacent rather than separated because an ortho substituted phenol

has an additional path three term contributed to $^{3}\chi_{p}^{v}$ relative to meta or para substitution. This arrangement predicts a higher toxicity based on eq. 2. There are no clear-cut cases in the data set to illustrate this situation. The three dimethyl compounds have toxicities which are rather close in terms of the experimental error. However, the utility of the $^{3}\chi_{p}^{v}$ index in describing toxicity in compounds which differ not so much in molecular size as in atom type and arrangement may nonetheless be demonstrated as follows. Consider the compounds in Table 1 which contain only halogens: compounds 1,2,13,14,20,21,23,24 (This set also includes phenol and 4-chloro-3-methylphenol in order to make a larger set.) Correlation with the $^{3}\chi_{p}^{v}$ index yields the following:

$$pLC50 = 0.95\ ^{3}\chi_{p}^{v} + 2.70$$

$$r = 0.95,\ s = 0.23,\ F = 51,\ n = 8$$

Comparison of correlation with number of atoms, $r = 0.63$, reveals the power of the $^{3}\chi_{p}^{v}$ index in this data set.

The molecular connectivity analysis of 25 substituted phenols leads to an equation which successfully models several key features of structure which influence the observed toxicity. Treatments of the equation reveal a very low probability that random correlations with the toxicity are present. Further, the equation is found to be quite stable in the presence of deletions from the database.

Direct structural interpretations of the equation are presented. In general an increase in molecular size, at least within the limits described by the database, results in increased toxicity. The presence of heteroatoms of the second quantum level lowers toxicity (relative to a methyl group) whereas atoms of higher quantum levels increase the toxicity. Adjacent ring substitution leads to higher toxicity relative to nonadjacent substitution.

The equation is considered to be sufficiently robust to permit prediction of other phenol toxicities.

ACKNOWLEDGMENTS We wish to acknowledge the support of this work by the U.S. Environmental Protection Agency under Cooperative Agreement CR-8-8923. We thank C.J. Cove, now at the Cornell University Medical School, and S.A. Henck at Eastern Nazarene College for their assistance in computer processing.

REFERENCES

Hall LH, Kier LB (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York

Hall LH, Kier LB (1978) A comparative analysis of molecular connectivity, Hansch, Free-Wilson and Darc-Pelco methods in the SAR of halogenated phenols. Eur. J. Med. Chem. 13:89-92

Hall LH, Kier LB (1981) The relation of molecular connectivity to molecular volume and biological activity. Eur. J. Med. Chem. 16:399-406

Hall LH, Kier LB (1983) An additivity model for the aquatic toxicity of substituted benzenes. Abstracts, Third Annual Meeting of the Society of Environmental Toxicology and Chemistry, Arlington, VA

Hall LH, Kier LB (1983) Structural information and a flexibility index from the molecular connectivity path-3 index. Quant. Struct. -Act. Rel. 2:55-59

Kier LB (1980) Molecular connectivity as a descriptor of structure for SAR analysis. In: Yalkowsky, SH, Physical chemical properties of drugs. Marcel Dekker, New York

Kier LB, Hall LH (1977) Structure-activity studies on hallucinogenic amphetamines using molecular connectivity. J. Med. Chem., 20:1631-1636

Kier LB, Hall LH (1981) Derivation and significance of valence molecular connectivity. J. Pharm. Sci. 70: 583-589

Kier LB, Hall LH (in press) SAR studies on the toxicities of benzene derivatives I. An additivity model. J. Environ. Tox. Chem.

Toplis J, Costello R (1972) Chance factors in studies on quantitative structure-activity relationships. J. Med. Chem. 15: 1066-1072

Received August 8, 1983; accepted October 24, 1983