REFERENCES

Adashek, K., Grossman M. I., (1963) Proc. Soc. Exp. Biol. Med. 112: 629–631

Leonards, J. R., Levy, G. (1969) J. Pharm. Sci. 58: 1277–1279

Leonards, J. R., Levy, G. (1972) Arch. Int. Med. 129: 457–460

Phillips, B. M. (1973) Toxicol. Appl. Pharmacol. 24: 182–189

Scheffe, H. (1967) The Analysis of Variance, Wiley & Sons, New York. p 73

Seegers, A. J. M., Jager, L. P. Van Noordwijk, J. (1978) J. Pharm. Pharmacol. 30: 84–87

Seegers, A. J. M., Jager, L. P. Van Noordwijk, J. (1979) Ibid. 31: 840–848

J. Pharm. Pharmacol. 1981, 33: 62–63
Communicated October 3, 1980

0022–3573/81/010062–02 $02.50/0

QSAR with random biological data

R. B. BARLOW, *Dept. of Pharmacology, Medical School, University of Bristol, Bristol BS8 1TD, U.K.*

In 1962 Hansch et al suggested that biological activity might be quantitatively related to chemical properties by a modification of the Hammett equation:

$$\log \frac{1}{C} = k\pi + k'\pi^2 + \rho\sigma + k''$$

where C is the concentration of a compound producing a standard biological response, σ is the Hammett substituent constant, expressing effects on electron distribution, and π is a parameter expressing effects on lipophilicity and is calculated from the change in log-partition coefficient between n-octanol and water. Hansch et al obtained values of σ and π for the substituents in 20 phenoxyacetic acids, measured the concentrations of the compounds which produced a standard growth of *Avena* coleoptiles and used the method of least squares to calculate the coefficients ρ, k, k′ and k″. Some idea of the goodness of fit was provided by comparing the experimental values of log 1/C with those calculated from the corresponding values of σ and π.

Subsequently many biological results have been fitted to equations of this type in attempts to establish quantitative structure-activity relationships (QSAR). The success of the operation is usually judged from the correlation coefficient, r, and standard deviation, s (Hansch & Fujita 1964; Tute 1971). For example, Hansch & Fujita reported that the phenol coefficients (PC) for 35 compounds tested against *M. pyogenes* var. *aureus* could be fitted to the equation

$$\log PC = 0{\cdot}001\pi^2 + 0{\cdot}953\pi - 0{\cdot}210\sigma + 0{\cdot}134$$

with r = 0·977 and s = 0·230. The correlation coefficient should represent

$$\sqrt{\frac{\text{explained variation}}{\text{total variation}}}$$

(Spiegel 1972) and the standard deviation is derived from the variance unexplained by regression,

$$S(PC_{observed} - PC_{calculated})^2$$

Because there are four coefficients (ρ, k, k′ and k″) to be calculated it might be expected that at least with small numbers of results some degree of correlation is inevitable (with only four results it should be perfect). Further, the limitations of the sensitivity of the method for measuring biological activity may tend to yield numbers which favour some degree of correlation. With a test capable of assessing activity over a million-fold range, the values of log 1/C will lie within 6 units but there is often great uncertainty with very weak compounds as to what the figure should be—a totally inactive compound cannot be represented on a log scale. It is therefore likely that an analysis will be restricted to those compounds whose activity lies within the range in which the test is considered to give reliable estimates; often this is less than 6 units. A further bias may also be introduced by the limited choice of substituents studied. Ideally the values of π and σ should be randomly distributed but this is seldom achieved and it is common to find that with both these parameters there are more positive values than negative ones.

This note describes an attempt to obtain some idea of the extent to which these factors may contribute to a correlation by examining some published results and replacing the experimental values of log 1/C by random values lying in the same range. Calculations were made with a Commodore PET 2001 computer and the procedure for obtaining the least-squares fit is outlined elsewhere (Barlow 1980). Studies were made with the original results involving 20 phenoxyacetic acids (Hansch et al 1962), with those for the 35 phenols referred to above (Hansch & Fujita 1964), and with results from the same paper for the toxicity of 14 benzoic acids to mosquito larvae. In each instance the original values for π and σ for each compound were taken and the computer's random number generator (RND(1)) was used to provide an 'estimate' of the biological activity. This function produces numbers between 0 and 1 and in a test of 20 000 values the distribution was as follows: <0·1, 1968; <0·2 but

>0·1, 2014; <0·3 but >0·2, 2000; <0·4 but >0·3, 1987; <0·5 but >0·4, 2000; <0·6 but >0·5, 1966; <0·7 but >0·6, 2010; <0·8 but >0·7, 2000; <0·9 but >0·8, 2085; <1·0 but >0·9, 1970. The mean was 0·502. Such numbers were used to produce values lying in the same range as the original estimates of log 1/C. The values of π, σ and the random-generated 'estimate' of biological activity were then fitted by the method of least-squares to the Hansch equation and the correlation coefficient was calculated. The process was repeated 800 times with each set and the values of r are shown in histograms in Fig. 1.

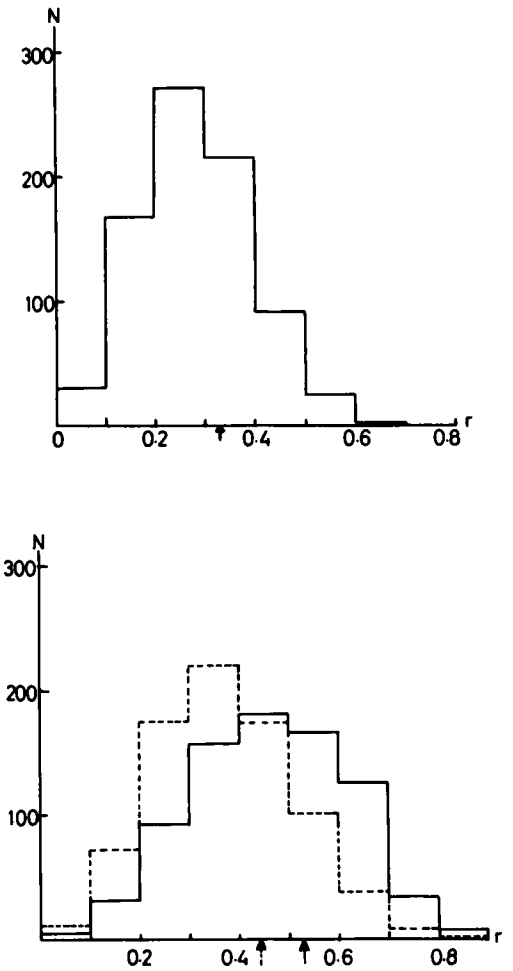

FIG. 1. Effect of group size on correlation coefficients obtained with random-generated biological activity. The histograms show the number of instances (N, out of 800) in which the correlation coefficient, r, lay within the range indicated. The upper section shows the results with values of π and σ for the 35 phenols and values of logPC in the range -1 to $+4$; the lower section shows (broken line) the results with values of π and σ for the 20 phenoxyacetic acids and values of log 1/C in the range 0 to 6·5 and (full line) with values of π and σ for the 14 benzoic acids and log 1/C in the range 1 to 5. The arrows indicate the values of r for significance at the level $P = 0{\cdot}05$.

With the π and σ values for the 35 phenols, values of r between 0·5 and 0·6 were obtained in 24 out of 800 instances (3%). With the 20 phenoxyacetic acids 4·75% of the values of r lay between 0·6 and 0·7 and 1% were better than 0·7. With the 14 benzoic acids 4·25 of the values of r were between 0·7 and 0·8 and 1% were actually better than 0·8. The corresponding values of $t\left(=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$ are 4·31, 4·16 and 4·62 for r = 0·6, 0·7, and 0·8 respectively and for the appropriate number of degrees of freedom these are all highly significant ($P < 0{\cdot}001$). For $P = 0{\cdot}05$ the limiting values of r for significance are 0·532 (n = 14), 0·444 (n = 20) and 0·327 (n = 35) and from the figure it can be seen that this was exceeded in over 20% of the random-generated results for benzoic acids and phenoxyacetic acids and with almost as high a proportion of the figures for the phenols.

It appears that with these particular results considerable bias is built in which favours correlation and the correlation coefficient is not by itself an adequate guide to the trust which can be placed in quantitative structure-activity relations of this type. It is possible that statistical analysis may indicate the extent to which the limited number of results, the limited range of values of log 1/C, and the limited choice of values of π and σ may each contribute to a spurious correlation. A direct assessment of their combined effects, however, can easily be obtained as described here, by replacing the experimental biological data by random-generated values lying in the same range and these should perhaps be included as a control, so that the extent of the improvement achieved with the experimental data can be assessed.

REFERENCES

Barlow, R. B. (1980) Quantitative Aspects of Chemical Pharmacology. Croom Helm, London, pp. 227–228

Hansch, C., Fujita, T. (1964) J. Am. Chem. Soc. 86: 1616–1626

Hansch, C., Maloney, P. P., Fujita, T., Muir, R. M. (1962) Nature (London) 194: 178–180

Spiegel, M. R. (1972) Statistics SI ed., McGraw-Hill, New York, p. 243

Tute, M. S. (1971) in: Harper, N. J., Simmonds, A. B. (eds) Advances in Drug Research, Vol. 6, Academic Press, London, pp 2–77