

Journal of Medicinal Chemistry

© Copyright 1964 by the American Chemical Society

VOLUME 7, NUMBER 4

JULY 6, 1964

A Mathematical Contribution to Structure-Activity Studies

SPENCER M. FREE, JR., AND JAMES W. WILSON

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pennsylvania

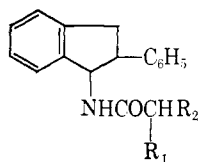
Received February 4, 1964

A mathematical technique is suggested as a means of describing structure-activity relationships of a series of chemical analogs. The data requirements included specific side chain arrangements and performance characteristics of all analogs tested. Two examples illustrate the use of the additive mathematical model where the performance characteristics are measures of biological activity. The results rank the structural changes per position by estimating the amount of biological response attributed to each change. The estimates are both positive and negative. Several uses for the mathematical solution are suggested.

Organic chemists who study analog series relate differences in structure to performance characteristics of each compound. The introduction of electronic computers offers opportunities to enhance this effort. Information retrieval techniques provide chemists with selected lists of analogs associated with biological response data. When such data have not been developed through first hand experience, and/or a large number of compounds are available, the determination of structure-activity relationships is more difficult. When the available data meet a limited number of restrictions, mathematical techniques can supplement the organic chemist's intuition.

This paper will develop some of the reasoning behind the proposed mathematical models. The models will then be used in two examples.

The Models.—The simplest problem would include the structure-activity of four analogs developed through two different changes at a single carbon in the molecule. For example, consider



where R_1 is H or CH_3 and R_2 is $\text{N}(\text{CH}_3)_2$ or $\text{N}(\text{C}_2\text{H}_5)_2$.

These are analgesic compounds where one biological response of interest is the LD_{50} . Results of the LD_{50} tests are as follows.

R_2	R_1		Average
	H	CH_3	
$\text{N}(\text{CH}_3)_2$	2.13	1.64	1.885
$\text{N}(\text{C}_2\text{H}_5)_2$	1.28	0.85	1.065
	1.705	1.245	1.475

This table shows that the analog with $R_1 = \text{H}$ and $R_2 = \text{N}(\text{CH}_3)_2$ has an LD_{50} of 2.13 mg./10 g. and is probably the best of the four. The statement is qualified because such tests are subject to biological variation.

The contribution of each substituent is easy to determine from these data. Comparing the average for the R_1 substituents, one can express the H contribution on the R_1 position as $1.705 - 1.475 = +0.23$ and the CH_3 contribution on the R_1 position as $1.245 - 1.475 = -0.23$. Here the symmetry ($+0.23$ and -0.23) is built into the solution. Likewise one can show for the R_2 substituent that the average contribution for the $\text{N}(\text{CH}_3)_2$ side chain is $+0.41$, so the other side chain contributes -0.41 to the average.

The example is presented to illustrate the additive property of the analogs. The mathematical models described in this paper will be based upon the assumption that there is some such "additivity" in series of analogs.

Continuing with the example to illustrate the mathematical model one writes a formula as follows

response = average + effect of R_1 substituent + effect of R_2 substituent

$$\text{LD}_{50} = \mu + a[\text{H}] + a[\text{CH}_3] + b[\text{N}(\text{CH}_3)_2] + b[\text{N}(\text{C}_2\text{H}_5)_2]$$

where μ = over-all average

$a[\text{H}]$ = contribution of H substituent at position R_1

$a[\text{CH}_3]$ = contribution of CH_3 substituent at position R_1

$b[\text{N}(\text{CH}_3)_2]$ = contribution of $\text{N}(\text{CH}_3)_2$ substituent at position R_2

$b[\text{N}(\text{C}_2\text{H}_5)_2]$ = contribution of $\text{N}(\text{C}_2\text{H}_5)_2$ substituent at position R_2

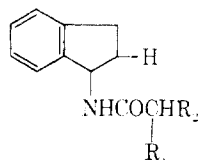
Substituting the actual LD_{50} values into this equation will produce four equations with five unknowns. However, the example shows

$$a[\text{H}] + a[\text{CH}_3] = 0; \text{ and } b[\text{N}(\text{CH}_3)_2] + b[\text{N}(\text{C}_2\text{H}_5)_2] = 0$$

or $a[\text{H}] = -a[\text{CH}_3]$; and $b[\text{N}(\text{CH}_3)_2] = -b[\text{N}(\text{C}_2\text{H}_5)_2]$

By imposing this restriction one need solve for only μ , $a[\text{H}]$, and $b[\text{N}(\text{CH}_3)_2]$. This reduces the mathematical problem to the solution of four equations and three unknowns. This is readily solved by least squares.¹

The fact that one needs only three equations for this problem suggests that one could estimate all the necessary unknowns if given data for only three of the four compounds. This is true on the assumption that the additive model holds. However, the problem is often complicated by poor precision of the biological activity measurements. What appears to be completely additive can be misleading just as what appears to have little or no additivity can be misleading. In the analogs



where R_2 is $\text{N}(\text{CH}_3)_2$ or $\text{N}(\text{C}_2\text{H}_5)_2$ and R_1 is H or CH_3 , the LD_{50} data are as follows.

R_2	R_1	
	H	CH_3
$\text{N}(\text{CH}_3)_2$	2.75	?
$\text{N}(\text{C}_2\text{H}_5)_2$	1.90	1.55

With or without the previous example one might predict that the missing datum would be a value larger than 2.00 based upon the additive model. Takahashi, *et al.*,² report a value of 1.77. Without repeated testing one does not know whether the disagreement is due to biological variation, nonadditivity, or both.

A more practical problem arises when additional variations in R_1 and R_2 are considered. Rarely are all the compounds of interest made and tested. Any structure-activity interpretation must cope with missing data as well as variation in the data available. Choosing the best compound can be difficult. Estimating the effect of each substituent is often ignored.

By assuming that some of the biological activity of the compounds is due to additivity, a mathematical model can be written to help describe this activity. If there are m variations in R_1 and n in R_2 there are $m \times n$ different compounds of interest which can be arranged in a two-way table. The mathematical model requires data for at least $m + n - 1$ compounds. A few more test results will be needed if the available data are arranged in an unusual manner in the table.

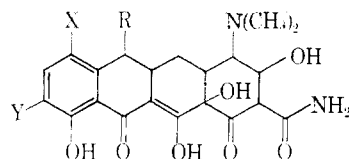
Additional substituents at other positions further complicate the structure-activity judgments. Again, by assuming that some of the order is based upon additivity one can write a mathematical model for the situation. Not all positions must be restricted to additivity. By isolating the additive portion, the solution can suggest where nonadditive effects may lie.

(1) R. L. Anderson and T. A. Bancroft, "Statistical Theory in Research," McGraw-Hill Book Company, Inc., New York, N. Y., 1952, p. 168.

(2) T. Takahashi, H. Fujimura, and K. Okamura, *J. Pharm. Soc. Japan*, **82**, 1597 (1962).

Examples.—Two examples are presented to illustrate the application of the mathematical models.

1.—Spencer, *et al.*,³ reported the *in vitro* inhibitory potencies against *Staphylococcus aureus* of ten disubstituted tetracyclines. Biological activity was expressed as a potency in comparison with tetracycline, which had an activity of 100 where a high biological activity is desirable. The analogs included the following



where R was H or CH_3 ; X was Br , Cl , or NO_2 ; and Y was NO_2 , NH_2 , or CH_3CONH .

The problem includes $2 \times 3 \times 3 = 18$ compounds of possible interest, of which only ten were made and tested. One can choose the best compound among the ten from the *in vitro* data. Yet, one would like some assurance that this choice is also the best among the 18 considered. Table I, showing the data, is expanded to illustrate the equations written for a model.

TABLE I
BIOLOGICAL ACTIVITY OF TEN TETRACYCLINES

Compound	Compound identification							Biological activity	
	H	CH_3	NO_2	Cl	Br	NO_2	NH_2		CH_3CONH
III	1		1			1			60
IV	1			1					21
V	1				1	1			15
VI	1			1			1		525
VII	1				1		1		320
VIII	1		1				1		275
IX		1	1				1		160
X			1					1	15
XI					1		1		140
XII		1			1			1	75

Following the rules set out (under "Models"), one writes a series of 10 equations in 6 unknowns. (There are really 9 unknowns which reduce to 6 because the contributions at each position sum to zero.) The results are presented in Table II.

TABLE II
CONTRIBUTION OF STRUCTURAL CHANGES^a


R	Side chain positions			Y	
	X				
$a[\text{H}]$	75	$b[\text{Cl}]$	84	$c[\text{NH}_2]$	123
$a[\text{CH}_3]$	-112	$b[\text{Br}]$	-16	$c[\text{CH}_3\text{CONH}-]$	18
		$b[\text{NO}_2]$	-26	$c[\text{NO}_2]$	-218

^a The solution includes these restrictions: $6a[\text{H}] + 4a[\text{CH}_3] = 0$; $2b[\text{Cl}] + 4b[\text{Br}] + 4b[\text{NO}_2] = 0$; and $5c[\text{NH}_2] + 2c[\text{CH}_3\text{CONH}-] + 3c[\text{NO}_2] = 0$. The over-all average was 161.

If one considers the total variation of the biological activity as 100%, this additive model accounted for 90.6% of the total. The results suggest that the best compound in the series would have $R = \text{H}$, $X = \text{Cl}$, and $Y = \text{NH}_2$. The estimated biological activity would be $161 + 75 + 84 + 123 = 443$. The actual response for this compound (VI) was 525. The dif-

(3) J. L. Spencer, J. J. Hlavka, J. Petisi, H. M. Krazinski, and J. H. Boothe, *J. Med. Chem.*, **6**, 405 (1963).

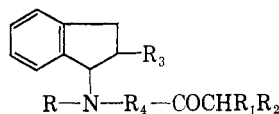
TABLE III
BIOLOGICAL ACTIVITY OF TWENTY-NINE INDANAMINES

Compd. no.	Compound identification													Biological activity	
	R		R ₁			R ₂			R ₃		R ₄		ED ₅₀	LD ₅₀	
	H	CH ₃	H	CH ₃	C ₂ H ₅	N(CH ₃) ₂	N(C ₂ H ₅) ₂		H	C ₆ H ₅	-	+			
19	1		1			1			1		1		0.46	2.75	
20	1		1				1		1		1		0.47	1.90	
21	1		1					1	1		1		0.30	5.00	
22	1			1		1			1		1		0.24	1.77	
23	1			1			1		1		1		0.48	1.55	
24	1			1				1	1		1		0.46	5.20	
25	1				1	1			1		1		0.22	1.75	
26	1				1		1		1		1		0.30	1.58	
27	1				1			1	1		1		0.46	5.00	
29		1	1					1	1		1		0.25	3.99	
30		1		1		1			1		1		0.21	1.28	
31		1		1		1			1		1		0.38	4.18	
32		1			1			1	1		1		0.38	3.02	
33	1		1							1	1		0.43	2.13	
34	1		1				1			1	1		0.32	1.28	
35	1		1			1		1	1	1	1		0.70	2.60	
36	1					1			1	1	1		0.37	1.64	
37	1					1	1		1	1	1		0.16	0.85	
38	1							1	1	1	1		0.53	3.95	
39	1				1					1	1		0.18	1.49	
47	1		1		1				1			1	0.21	2.70	
48	1		1				1		1			1	0.21	5.10	
49	1		1					1	1		1		0.52	7.00	
50	1			1		1			1		1		0.46	2.60	
51	1			1			1		1		1		0.36	3.57	
52	1			1				1	1		1		0.41	5.00	
53	1				1	1			1		1		0.37	6.20	
54	1				1		1		1		1		0.46	3.00	
55	1				1			1	1		1		0.38	7.40	


ference between estimated and observed activity would be due to biological variation in the *in vitro* procedure and/or a nonadditive component in the analog series. Since the model accounts for 90% of the total variation it is unlikely that there is a nonadditive component in this series.

Table II suggests that one compound (VI) is superior in the series. However, no effort is made here to emphasize the ability to estimate the activity of any specific compound. Experience to date suggests that the ranking of the potential functional groups at each position has the most utility.

2.—Takahashi, *et al.*,² reported on analgesic indanamine derivatives. The data included two criteria of biological activity. Both ED₅₀ and LD₅₀ data were expressed as mg./10 g. upon intraperitoneal administration to mice. The authors also expressed the therapeutic index LD₅₀/ED₅₀. Mathematical models were fitted to the original criteria because ratios are quite variable. The analogs included the following compounds.




R = H or CH₃; R₁ = H, CH₃, or C₂H₅;

R₂ = N(CH₃)₂, N(C₂H₅)₂, or ;


R₃ = H or C₆H₅; R₄ = nothing or —CONH—

The R₄ substitution is used here to show how one might include such compounds in the mathematical problem. The problem includes 2 × 3 × 3 × 2 × 2 = 72 compounds of interest. Table III shows that of these, **72** and **29** were tested in both biological tests. Following the rules for a least-squares solution one writes 29 equations in 8 unknowns. The remaining 29 - 8 = 21 equations provide estimates of biological variation and can be used to estimate inadequacies of the simple additive model. Here one develops two solutions for the two biological phenomena. The results of one mathematical solution are presented in Table IV.

The additive model accounted for 81% of the variation in the LD₅₀. Table IV nominates two compounds as least toxic in the series. These have R = H, R₂ = , R₃ = H, R₄ = —CONH—, and R₁ = C₂H₅ or H. Both compounds were tested (numbers **49** and **55**). The solution cannot be extended to state a preference. The solution also suggests that the results for compound **53** may have been influenced by extreme biological variation, although this could also be due to a nonadditive effect.

The same model accounted for only 31% of the variation in the ED₅₀ data. This was not enough to consider the model useful for nominating the better compounds in the series. Several other models were tried and all failed. The results suggested that, although the range of ED₅₀ values was more than four-fold (0.16 to 0.70), most of the differences could be due to the variation in the test procedure.

TABLE IV
 ESTIMATED CONTRIBUTIONS OF ALL SUBSTITUENTS^a

<i>n</i> ^b	R	LD ₅₀	<i>n</i>	R ₁	LD ₅₀	<i>n</i>	R ₂	LD ₅₀	<i>n</i>	R ₃	LD ₅₀	<i>n</i>	R ₄	LD ₅₀
25	a[H]	0.12	8	b[C ₂ H ₅]	0.17	11	c[	1.60	22	d[H]	0.20	9	e[+]	1.23
4	a[CH ₃]	-0.76	10	b[H]	0.14	10	c[N(CH ₃) ₂]	-0.82	7	d[C ₆ H ₅]	-0.62	20	e[-]	-0.55
			11	b[CH ₃]	-0.25	8	c[N(C ₂ H ₅) ₂]	-1.17						

^a The over-all LD₅₀ average was 3.29. ^b *n* = number of compounds with the substituents.

 TABLE V
 EXPANDED INDANAMINES BIOLOGICAL ACTIVITY IDENTIFICATION











Compt. no.	R ₁ × R ₂ side chains												R ₃	R ₄		LD ₅₀	LD ₉₅
	H	CH ₃	N(CH ₃) ₂	N(C ₂ H ₅) ₂								H		C ₆ H ₅	-		
19	1		1									1		1		0.46	2.75
20	1			1								1		1		0.47	1.90
21	1				1							1		1		0.30	5.00
22	1					1						1		1		0.24	1.77
23	1						1					1		1		0.48	1.55
24	1							1				1		1		0.46	5.20
25	1								1			1		1		0.22	1.75
26	1									1		1		1		0.30	1.58
27	1										1	1		1		0.46	5.00
29		1			1							1		1		0.25	3.99
30		1				3						1		1		0.21	1.28
31		1					1					1		1		0.38	4.18
32		1									1	1		1		0.38	3.02
33	1		1									1	1			0.43	2.13
34	1			1								1	1			0.32	1.28
35	1				1							1	1			0.70	2.60
36	1					1						1	1			0.37	1.64
37	1						1					1	1			0.35	0.85
38	1							1				1	1			0.53	3.95
39	1								1			1	1			0.18	1.19
47	1		1									1		1	0.21	2.70	
48	1			1								1		1	0.21	5.19	
49	1				1							1		1	0.52	7.00	
50	1					1						1		1	0.46	2.60	
51	1						1					1		1	0.36	3.57	
52	1							1				1		1	0.41	5.00	
53	1								1			1		1	0.37	6.20	
54	1									1		1		1	0.46	3.00	
55	1										1	1		1	0.38	7.40	

 TABLE VI
 ESTIMATED SIDE-CHAIN CONTRIBUTIONS FOR EXPANDED MODEL

<i>n</i> ^a	R	LD ₅₀	<i>n</i>	R ₁	R ₂	LD ₅₀	<i>n</i>	R ₃	LD ₅₀	<i>n</i>	R ₄	LD ₅₀
25	H	0.12	3	C ₂ H ₅		1.75	22	H	0.22	9	+	1.23
4	CH ₃	-0.73	4	H		1.56	7	C ₆ H ₅	-0.69	20	-	-0.55
			4	CH ₃		1.50						
			3	C ₂ H ₅	N(CH ₃) ₂	-0.22						
			3	H	N(C ₂ H ₅) ₂	-0.61						
			3	H	N(CH ₃) ₂	-0.84						
			1	CH ₃	N(CH ₃) ₂	-1.26						
			3	CH ₃	N(C ₂ H ₅) ₂	-1.38						
			2	C ₂ H ₅	N(C ₂ H ₅) ₂	-1.68						

^a *n* = the number of compounds with the side chain.

To demonstrate how one can expand the concept of the additive model to include more specific changes, one of the solutions used as an attempt to explain the ED₅₀ data is illustrated in Table V. This model accounts for all nine changes in the R₁ and R₂ positions. The mathematical model is still 29 equations, this time in 12 unknowns. The ED₅₀ solution accounted for only 36% of the variation, so the estimated contributions are not included in the summary table. The LD₅₀ model accounted for 80+% of the variation, and the results are shown in Table VI. The results from

Tables IV and VI are comparable, suggesting that the more complex model added nothing new to the solution.

Discussion

This mathematical model approach is suggested as one way to summarize data generated for a series of chemical analogs. There should be times when the solution will supplement the reasoning of the organic chemist. This quantitative approach may also pro-

vide a means of transferring information from one series to another.

This scheme should not be looked upon as a means for identifying the best single compound. Data from larger series of analogs would be expected to suggest a few desirable substituents at more than one position. The rank order of related substituents within a position would be expected to have meaning. Some solutions should suggest untried substituents as good leads.

The proposed models should not be criticized as ignoring the combination of several substituents that produces a biological response far in excess of the additive estimation. Such results will appear in some analog series. Such situations might be identified by a

graph of the individual differences of "estimated response minus actual response" for all compounds.

Successful solutions can provide reasonable estimates of inherent variation within the testing system. These may not otherwise be available without repeated testing of the same compounds. Solutions that fail can suggest that the substituents may not be altering the desirable performance characteristics of the analogs.

The suggested mathematical models do not compensate for the three dimensionality of compounds, pH, pK_a , or other similar physical properties. Perhaps, in time, these can be built into the models for better estimation.

The Metabolic Fate of Thiabendazole in Sheep¹

DOMINICK J. TOCCO,

Merck Institute for Therapeutic Research, Rahway, New Jersey

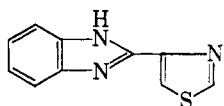
RUDOLF P. BUHS, HORACE D. BROWN, ALEXANDER R. MATZUK, HOLLY E. MERTEL,
ROBERT E. HARMAN, AND NELSON R. TRENNER

Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey

Received September 24, 1963

The synthesis of C¹⁴- or S³⁵-labeled thiabendazole is described. Following oral administration of this anthelmintic to sheep (50 mg./kg. of body weight), the physiological disposition and metabolic fate of the compound have been investigated. The animals were sacrificed from 6 hr. to 30 days after dosage and the distribution of the drug was studied in urine, feces, plasma, and tissues using liquid scintillation or gas-flow counting. Sheep excrete approximately 75% of the dose in the urine and 14% in the feces in 96 hr. Although thiabendazole is distributed throughout most tissues of the body, only fractional parts per million were detectable in tissue after a few days. The major metabolites were isolated from urine and identified as 5-hydroxythiabendazole which exists either free or conjugated as the glucuronide or sulfate.

Thiabendazole [2-(4'-thiazolyl)benzimidazole], a compound having the following chemical structure



is a new and highly effective drug used in the treatment of helminthiases. The compound has a broad anthelmintic spectrum affecting numerous gastrointestinal roundworms and certain tapeworms.^{2,3} More recently the effectiveness of this drug on trichinosis in mice, rats,⁴ and swine⁵ has been reported.

The present report concerns itself with the absorption, excretion, metabolic transformation, tissue distribution, and retention of thiabendazole in sheep. Radioisotopically labeled drug has been utilized as a guide in the isolation of the various metabolites and to follow its physiological disposition. It is shown that the drug is metabolized in part to a compound which is

hydroxylated in the benzimidazole ring. Further metabolism of this hydroxylated product results in the formation of its glucuronide and sulfate ester.

Experimental

Materials and Methods. Synthesis of C¹⁴-Labeled Thiabendazole.—Starting material for the synthesis of thiabendazole labeled with C¹⁴ in the benzene ring portion of the molecule was uniformly ring-labeled aniline⁶ (I). Aniline hydrochloride reacted smoothly with oxalyl chloride in boiling benzene to give oxanilide⁷ (II). Using a modification of a procedure disclosed in the patent literature,⁸ the oxanilide was sulfonated, nitrated, and hydrolyzed to give crude *o*-nitroaniline (III). After purifying the *o*-nitroaniline by crystallization and sublimation, this intermediate was caused to react with 4-thiazolecarbonyl chloride to give the corresponding nitroanilide⁹ (IV). Catalytic reduction of the *o*-nitroanilide gave the corresponding aminoanilide (V), which, upon refluxing with acid, cyclized to the hydrochloride of thiabendazole. The free base, thiabendazole (benzene ring carbon-14) (VI), was liberated by treatment of the hydrochloride in water with sodium bicarbonate.

Oxanilide Ring C¹⁴ (II).—To 76 ml. of azeotropically dried benzene was added 5.30 g. (0.04 mole) of I (30.0 mc. of C¹⁴)

(1) A preliminary report was presented before the Federation of American Societies for Experimental Biology, Atlantic City, N. J., April 14–18, 1962.

(2) H. D. Brown, A. R. Matzuk, I. R. Ilves, L. H. Peterson, S. A. Harris, L. H. Sarett, J. R. Egerton, J. J. Yakstis, W. C. Campbell, and A. C. Cuckler, *J. Am. Chem. Soc.*, **83**, 1764 (1961).

(3) A. C. Cuckler, *J. Parasitol.*, **47**, 37 (1961).

(4) W. C. Campbell, *ibid.*, **47**, 37 (1961).

(5) W. C. Campbell and A. C. Cuckler, *Proc. Soc. Exptl. Biol. Med.*, **110**, 124 (1962).

(6) Aniline-C¹⁴ (uniform) was supplied by Merck & Co., Limited, Montreal, Canada.

(7) M. J. Th. Bornwater, *Rec. Trav. Chim.*, **31**, 105 (1912).

(8) Anilinoel-Fabrik A. Wulfung, Elberfeld, German Patent 65,212; Friedl., III, 44 (1894).

(9) The procedure for the sequence of reactions from *o*-nitroaniline to thiabendazole was furnished by E. E. Harris and R. B. Currie of the Chemical Division of Merck & Co., Inc.