

**Acknowledgment.** The authors wish to thank Professor Tamura of the Faculty of Pharmaceutical Sciences, Osaka University, for his advice on methods of synthesis.

## References

- (1) (a) E. J. Ariens, *Ann. N. Y. Acad. Sci.*, **139**, 606 (1967); (b) A. F. Crowther and L. H. Smith, *J. Med. Chem.*, **11**, 1009 (1968); (c) A. F. Crowther, K. J. Lorghlin, L. H. Smith, R. W. Turner, and T. M. Wood, *ibid.*, **12**, 638 (1969); (d) R. Howe, *ibid.*, **12**, 642 (1969); (e) J. F. Giudicelli, H. Schmitt, and J. R. Boissier, *J. Pharmacol. Exp. Ther.*, **168**, 116 (1969); (f) B. Levy and M. Wasserman, *Brit. J. Pharmacol.*, **39**, 139 (1970); (g) R. P. Robson and H. R. Kaplan, *J. Pharmacol. Exp. Ther.*, **175**, 157 (1970); (h) C. F. Schwender, S. Farber, C. Blaum, and J. Schavel, *J. Med. Chem.*, **13**, 684 (1970); (i) M. Nakanishi, T. Muro, Y. Chihara, H. Inamura, and T. Naka, *ibid.*, **15**, 45 (1972); (j) K. Murase, K. Niigata, T. Mase, and M. Murakami, *Yakugaku Zasshi*, **92**, 1358 (1972); (k) Y. Kobayashi, T. Nakagaki, T. Oshima, S. Kumakura, K. Nakayama, and H. Koide, *Chem. Pharm. Bull.*, **20**, 905 (1972).
- (2) (a) Y. Tamura, M. Terashima, Y. Higuchi, and K. Ozaki, *Chem. Ind. (London)*, 1935, (1970); (b) M. Mano and Y. Tamura, Japan Patent 39,694 and 39,695 (1971).
- (3) (a) *Chem. Abstr.*, **50**, 4964e (1965); (b) F. Mayer, L. van Zutphen, and H. Philips, *Ber.*, **60**, 858 (1927); (c) J. D. London and J. Ogg, *J. Chem. Soc.*, 739 (1955); (d) N. Shigematsu, *Chem. Pharm. Bull.*, **9**, 970 (1961).
- (4) K. Ozaki and Y. Tamura, Japan Patent 38,789 (1971).
- (5) R. D. Robson and H. R. Kaplan, *J. Pharmacol. Exp. Ther.*, **175**, 157 (1970).

## Substructural Analysis. A Novel Approach to the Problem of Drug Design

Richard D. Cramer III,\* George Redl, and Charles E. Berkoff

*Technology Assessment, Smith Kline & French Laboratories, Philadelphia, Pennsylvania 19101. Received September 10, 1973*

Of the many approaches to the problem of drug design, those of greatest current utility and application are the regression techniques commonly associated with the names of Hansch<sup>1</sup> and Free-Wilson.<sup>2</sup> A severe limitation shared by these methods is their restriction to structurally closely related series of compounds. Thus they are inappropriate for (a) correlation of data where compounds fall into many different structural series or into no series at all; (b) prediction of active compounds outside a structural class of established biological interest. A second major limitation of these methods is their weakness in accommodating data represented by inactive compounds. In essence, existing structure-activity correlation methodologies are useful only for optimizing a previously recognized "lead" structure and not in generating new "leads."

The continuing need for drug design techniques that would be applicable to a broader range of problems led us to consider the mental model on which the medicinal chemist bases his search for new lead structures. An evident truth forming the basis of this model is that the biological activity of a molecule, or for that matter any other of its properties, must be accounted for by a combination of contributions from its structural components (substructures) and their intra- and intermolecular interactions. The very large body of information generated by even the most modest of screening programs requires the medicinal chemist to make additional simplifying assumptions, such as (a) the probability of a given biological activity can be usefully approximated by a first-order analysis of substructural contributions (*i.e.*, one ignoring interactions); (b) the contribution of a given substructure to the probability of activity can be obtained from data on previously tested compounds containing that substructure. The spe-

cific question we sought to answer empirically was whether a significant correlation could be obtained by systematically organizing existing sets of biological and substructural data to correspond with this mental model. (A previous approach to this problem using the statistical technique of cluster analysis appeared to show promise.<sup>3</sup>)

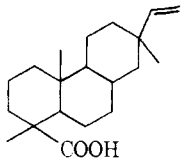
Existing schemes for the codification of substructures have been created solely in response to a need for selective retrieval of compounds from large files.<sup>4</sup> Most of the substructures that chemists habitually perceive are far more complex than the several-atom "fragments" of these codes. These limitations clearly applied even to the relatively rich "SK&F fragment code," which recognizes some 1200 fragments comprising functional groups, rings, chains, inorganic moieties, and 110 rather diffusely defined fragment combinations.<sup>5</sup> For our pilot study we nevertheless attempted to use this code for the analysis of the most structurally diverse testing experience available to us, consisting of 850 compounds examined for their antiarthritic-immunoregulatory effects in an adjuvant-induced rat model.<sup>6</sup> To remove inherent sample bias and to ensure that our analyses would not simply regenerate known information, compounds which were members of already recognized "lead" series were eliminated, leaving 770 compounds. Of these, 189 (24.5%) were active, producing a statistically significant reduction in hind paw volume during the secondary phase of the induced disease process. It should be noted that, since such activity is displayed by agents having quite varied pharmacological properties, neither the available biological data nor the SK&F fragment code were totally appropriate for our objective.

The first step was to prepare a substructure "experience table" summarizing the data. A "Substructure Activity Frequency" (SAF), defined for each substructure as (A/T), the ratio of the number of active compounds (A) containing that substructure to the number of tested compounds (T) containing the substructure, represents the contribution which that substructure can make to the probability of a compound being active. The experience table contained 492 SAF's corresponding to the complete set of 492 substructures (fragments) previously recognized and coded in the tested compounds.

We then computed for each compound a "Mean Substructure Activity Frequency" (MSAF), the arithmetic mean of the SAF values of the substructures present in that compound. A sample MSAF computation appears in Table I. The 770 compounds next were ranked by descending MSAF value. Since a meaningful correlation would be reflected in a tendency for compounds of higher MSAF value to be active more frequently, the 770 ranked compounds were partitioned into ten sets, each containing 77 compounds. Those sets with high MSAF values were indeed found to be active far more frequently than those with low MSAF values (Table II).

However, analysis of some individual MSAF computations showed that MSAF values could be strongly influenced by SAF values for substructures that were poorly represented within the total set of tested compounds. For example, the SAF for a unique fragment must take either of the extreme values of 1.0 or 0.0, depending on whether the compound in which it occurred was active or not, and the MSAF for that compound would thus be biased in a direction which would improve the apparent correlation. Thus, even though unique fragments contribute less than 1% of the quantity of information, their impact on the overall analysis is substantial. To remove this type of bias, and to estimate the predictive value of the method, we devised a novel computational approach.

Groups of ten compounds, selected at random, became

**Table I.** Sample Calculation of a Mean Substructure Activity Frequency (MSAF) for a Compound Which Contains a Substructure Unique among the 771 Compounds Tested for Antiarthritic-Immunoregulatory Activity


SK&F 37422 active

Fragment and description	Tested (T)	Active (A)	Substructure act. frequency (SAF)
HC2, aliphatic COOH	108	23	0.212
NL3, phenanthrene ring system, one unsaturation	1	1	1.000
000, carbon chain as functional group	198	49	0.247
003, methyl chain	309	83	0.268
006, quaternary carbon	78	18	0.230
018, unsaturated 2-carbon side chain	3	1	0.333
69/0, three rings	174	45	0.258
69/4, angular substitution	25	7	0.280
70/7, condensed alicyclic ring system	31	12	0.387
74/Y, multiple substituents	452	115	0.254
74/X, substituents $\alpha$ to fusion	163	45	0.276
74/0, substituents $\beta$ to fusion	160	47	0.293
74/5, 1,3 substituent pattern	100	33	0.330
75/X, geminal substitution	98	16	0.163
76/3, ring -C=	169	40	0.236
76/5, ring -CC=	111	23	0.207
		16	4.974
			0.3109 <sup>a</sup>

<sup>a</sup>MSAF = average of SAF's or 0.3109.

**Table II.** Occurrence of Antiarthritic-Immunoregulatory Activity among 770 Compounds of Varied Structure Ordered by Descending MSAF Value (Derived by Considering Contributions from All Tested Compounds) and Then Partitioned into Ten Sets

Set <sup>a</sup>	Av MSAF value within set	No. of active compds within set <sup>b</sup>
1	0.320	56
2	0.269	49
3	0.257	26
4	0.249	24
5	0.243	12
6	0.237	7
7	0.231	7
8	0.225	2
9	0.215	2
10	0.186	4

<sup>a</sup>Each set contains exactly 77 compounds. <sup>b</sup>The differences of these values from the random or mean value (18.9) are highly significant ( $p < 0.01$ ). <sup>c</sup>The differences of these two values from their mean (94.5) are highly significant ( $p < 0.01$ ).

the subjects of a simulation study for the prediction of activity. For each of these groups new SAF values were computed, including only the substructure-activity data obtained from the other 760 compounds. New MSAF values subsequently derived were used to rank the ten compounds within a group by descending MSAF order. This sequence of computations was repeated 77 times to include all of the 770 compounds, while a record was kept of the number of occasions that compounds in first rank,

**Table III.** Occurrence of Antiarthritic-Immunoregulatory Activity among 770 Compounds of Varied Structure When Ranked, within Groups of Ten, by "Predictive" MSAF Value (Derived by Excluding the Contributions of All Members of a Group to the Component SAF Values)

Rank <sup>a</sup>	Av MSAF value within set	No. of active compds within rank <sup>b</sup>
1	0.293	23
2	0.266	26
3	0.256	25
4	0.249	21
5	0.241	17
6	0.239	13
7	0.234	17
8	0.221	10
9	0.212	19
10	0.195	18

<sup>a</sup>Each rank contains exactly 77 compounds. <sup>b</sup>The differences of these values from the random or mean value (18.9) are not significant ( $p > 0.10$ ). <sup>c</sup>The differences of these two values from their mean (94.5) are significant ( $p < 0.02$ ).

second rank, and so on through tenth rank, within a group, were actually found to be active. By preventing a compound's actual test result from influencing either its own "predictive" MSAF value or the values of any other members of its group, this computational approach closely simulates approaches requiring the testing of additional compounds.

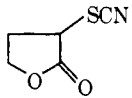
The results of this computation are expressed in Table III as the total number of active compounds occurring among each of the ten ranks. As would be expected, the association between ranked MSAF values and activity is weaker when the MSAF value of a compound cannot be influenced by its own test result. Systematic exclusion of "poorly represented" compounds (*i.e.*, those having fragments with low individual or low average *T* values) did not improve the predictive power of the correlation. Yet, since a difference in distribution of activity between the first five ranks and the second five ranks of the magnitude observed has less than a 2% probability of occurrence by chance, we reached the preliminary conclusion that substructural analysis is capable of predicting differential probabilities of activity in a set of "untested" structurally diverse compounds.

During the course of the above-described study, a further structurally diverse set of 703 compounds, not members of any "lead" series, had been tested for antiarthritic-immunoregulatory activity. To put the technique of substructural analysis to a practical test, we attempted to predict the distribution of activity within this new set using an experience table derived only from the initial analysis of 770 compounds. MSAF values for each of the new compounds were computed as before (Table I), except that in order to avoid the strongly biasing influence of unrepresentative SAF values substructures that had appeared in fewer than six of the original 771 compounds were excluded. Table IV shows the MSAF computation for a structure containing one such poorly represented substructure.

Furthermore, where SAF values for more than 10% of the substructures present in a given compound were unavailable, or unrepresentative in the above sense, the compound as a whole was regarded as insufficiently described by existing testing experience and was thus excluded from the predictive analysis. The fraction of compounds thus excluded amounted to less than one-third of the total set.

The distribution of the fraction of active compounds according to the value of the calculated MSAF parameter

**Table IV.** Sample Calculation of Mean Substructure Activity Frequency (MSAF) for One of 489 Compounds Whose Probability of Antiarthritic-Immunoregulatory Activity Was Predicted



SK&F 43248 inactive

Substructure and description	Occurrences in first 770 (T)	Occurrences in first 770 (A)	Frequency
			(SAF = A/T)
H26, lactone	51	8	0.1568
IWB, saturated C <sub>4</sub> O ring	68	18	0.2647
000, carbon isolated in a functional group	198	49	0.2474
862, Het-SCN	1	0	0.0000 <sup>a</sup>
69/Y, one ring	549	126	0.2295
71/Y, isolated heterocycle	343	77	0.2244
71/5, five-membered heterocyclic ring	120	32	0.2666
72/4, one oxygen in ring	213	56	0.2629
72/7, only one heteroatom in ring	328	83	0.2530
74/1, substitution $\alpha$ to a heteroatom	468	117	0.2500
74/2, substitution $\beta$ to a heteroatom	245	58	0.2367
74/4, 1,2 substitution	150	34	0.2266
77/6, X-C-C-C-Y	96	20	0.2083
78/Y, X-C-C=Y	113	22	0.1946
		13	3.0215
			MSAF = 0.2324

<sup>a</sup>The SAF value for substructure 862 is not included in the MSAF computation because of insufficient representation among the first 770 compounds. (See text.)

**Table V.** Occurrence of Antiarthritic-Immunoregulatory Activity among 489 Compounds Grouped According to Mean Substructure Activity Frequency (MSAF). MSAF Values are Calculated on the Basis of Data from 770 Previously Tested Compounds

MSAF range	No. tested	No. active <sup>a</sup>	Frequency of act.
>0.26	80	18	0.225
0.25-0.26	85	18	0.212
0.24-0.25	127	13	0.102
0.23-0.24	116	14	0.121
<0.23	81	11	0.136
Totals	489	74	

<sup>a</sup>The difference of these values from the random values (12.1, 12.9, 19.2, 17.6, and 13.0, respectively) is marginally significant ( $p < 0.1$ ). If the compounds with MSAF > 0.25 (top two MSAF ranges together) are compared to the compounds with MSAF < 0.25 (bottom three MSAF ranges together), the difference between the observed numbers of actives (36 and 38) and the most probable or "random" values (25 and 49, respectively) is highly significant ( $p < 0.01$ ).

ranges is shown in Table V. While there is no difference among the lower three MSAF ranges shown, activity is clearly and significantly ( $p < 0.01$ ) less frequent among compounds in the lower three ranges when compared with the higher two MSAF ranges. Considering the limitations of the biological data used and the coarse discriminatory power of the available substructural system, we are encouraged by the results.

The application of this method to other substructural systems (such as those based on Wiswesser notation) and to other sets of biological data is clearly indicated. In the context of large screening programs this technique may be of practical value, even in its present primitive form, by improving the efficiency of "lead" generation.

Although alternative computational procedures can be explored, more useful correlations (*i.e.*, prediction of larger differential probabilities of activity) will probably require a more sophisticated substructural system based on direct computer manipulation of complete structural records.<sup>7</sup> Ultimately, perhaps in an evolutionary process guided by substructural analysis of many biological data, this system could place emphasis on those molecular features that prove to be of fundamental significance to biological mechanisms. Advances in computers and programming technology<sup>8</sup> are beginning to make feasible the systematic study of factors as complex as three-dimensional structure, polarizability, bonded and nonbonded interactions, and solvation phenomena.

**Acknowledgment.** We thank Dr. A. D. Bender for stimulating discussion and continuing encouragement.

### References

- (1) C. Hansch, *Accounts Chem. Res.*, **2**, 232 (1969).
- (2) S. M. Free and J. W. Wilson, *J. Med. Chem.*, **7**, 395 (1964).
- (3) P. J. Harrison, *J. Appl. Stat.*, **17**, 226 (1968).
- (4) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structural Information," American Elsevier, New York, N. Y., 1972, pp 67-95.
- (5) P. N. Craig and H. M. Ebert, *J. Chem. Doc.*, **9**, 141 (1969).
- (6) D. T. Walz, M. J. DiMartino, and A. Misher, *J. Pharmacol. Exp. Ther.*, **178**, 223 (1971), and references cited therein.
- (7) M. Milne, D. Lefkowitz, H. Hill, and R. Powers, *J. Chem. Doc.*, **12**, 183 (1972); M. A. T. Rogers, *Chem. Ind. (London)*, 952 (1970).
- (8) W. T. Wipke, P. Gund, J. G. Verbalis, and T. M. Dyatt, Abstracts, 162nd National Meeting of the American Chemical Society, Washington, D. C., Sept 1971; R. J. Feldmann, S. R. Heller, and C. R. T. Brown, *J. Chem. Doc.*, **12**, 234 (1972); E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe, *J. Amer. Chem. Soc.*, **94**, 431, 440 (1972).

### Preparation and Antitumor Activity of 4'-Thio Analogs of 2,2'-Anhydro-1- $\beta$ -D-arabinofuranosylcytosine

N. Ototani and Roy L. Whistler\*

*Department of Biochemistry, Purdue University, Lafayette, Indiana 47907. Received November 1, 1973*

Among nucleosides with antitumor activity, 1- $\beta$ -D-arabinofuranosylcytosine has well-known activity against rodent and human neoplasms.<sup>1</sup> The drug has been used clinically against acute leukemia and lymphoma.<sup>2</sup> Because the compound produces megaloblastosis and chromosomal alteration in bone marrow,<sup>2a,3</sup> we have prepared a sulfur analog with the hope that it might be less toxic. Hopefully, also, the analog might be less rapidly deaminated to inactive spongouridine.<sup>4</sup>

Since 2,2'-anhydro-1- $\beta$ -D-arabinofuranosylcytosine is less easily deaminated<sup>5</sup> than the straight nucleoside, we have also prepared the 4'-thio analog of the anhydro nucleoside.

It is often observed that low yields of nucleosides are obtained by condensing 4-thio-D-ribofuranosyl derivatives with pyrimidine bases. However, the condensation of 1,2,3,5-tetra-*O*-acetyl-4-thio-D-ribofuranose and bis(trimethylsilyl)-*N*-acetylcytosine with stannic chloride as catalyst<sup>6</sup> gave a 65% yield of the acetylated  $\beta$ -D nucleoside I with the  $\alpha$ -D nucleoside in 3% yield.