

A Study of Structure-Activity Relationships of a Series of Diphenylaminopropanols by Factor Analysis†

Myra L. Weiner

Research Institute of Pharmaceutical Sciences, School of Pharmacy, University of Mississippi, University, Mississippi 38677

and Paul H. Weiner*

Department of Chemistry, University of Mississippi, University, Mississippi 38677. Received August 8, 1972

The mathematical technique of factor analysis was applied to the biological activity data of Keasling and Moffett. Using a data matrix consisting of the measured effects of 16 diphenylaminopropanols on 11 biological activity tests in mice, we found that 8 factors were required to span the data set such that all data were reproduced within ± 0.1 ln unit. Attempts were made to rotate physically significant structural drug parameters into the abstract factors of the space. Using the concept of a "uniqueness test," discussion will also be given concerning the interrelationships between the 11 biological activities reported.

Medicinal chemists have long been interested in utilizing drug structure-activity relationships in the design of new drugs. The technique of regression analysis has become a valuable tool in analyzing drug structure-activity relationships in a quantitative manner.¹ However, analysis of drug data by this approach is limited to a single biological activity. Since the usefulness of a new drug will depend upon its actions on many physiological parameters, as well as its absorption, distribution, and excretion in the intact animal, it would be helpful to study drug structure-activity relationships simultaneously on a wide gamut of biological tests.

Recently the mathematical technique of factor analysis² (F/A) has gained prominence as a tool in analyzing complex multidimensional problems in chemistry, such as solvent effects in nuclear magnetic resonance,³⁻⁶ activity coefficients in gas-liquid chromatography,⁷ structure-retention index data in gas chromatography,⁸⁻¹² and also in medical problems, such as prognosis of diseases.¹³ F/A can be applied to any problem in which the quantity being analyzed can be expressed as a linear sum of terms in product function form.⁴ Mathematically, this means that the data being factor analyzed must be expressible by an equation of the following form, namely

$$P(i,\alpha) = \sum_{j=1}^m U(i,j) V(j,\alpha) \quad (1)$$

Here $P(i,\alpha)$ is some measured property of the system of case i on variable α ; $U(i,j)$ is the j th case factor for the i th case; and $V(j,\alpha)$ is the j th variable factor for the α th variable, the sum being taken over the j important factors in the problem. Hansch^{1a} points out that $\log 1/\text{LD}_{50}$, lethal dose-50 in milligrams per kilogram, or $1/\text{ED}_{50}$, effective dose-50 in milligrams per kilogram, can be linearly related by regression analysis to various properties of the drugs. Therefore, we have applied factor analysis to the $\log \text{LD}_{50}$ or $\log \text{ED}_{50}$ data to linearize the resulting functional equations. One can factor analyze either $\log \text{LD}_{50}$ or $\log \text{ED}_{50}$ or the inverse of these quantities since the two results only differ by a sign change. We have chosen to apply our analysis to the former form. For the drug structure-biological activity problem, a solution in the following form must exist.

$$\log [\text{BA}(i,\alpha)] = \sum_j D(i,j) \cdot H(j,\alpha) \quad (2)$$

$\text{BA}(i,\alpha)$ is the α th measured biological activity of drug i ; $D(i,j)$ is the j th physical or chemical property of the i th drug; and $H(j,\alpha)$ is the j th physiological host parameter of the α th measured biological activity. The sum is taken over the j independent parameters needed to account for the data. Since regression analysis requires the same constraints^{1a} as F/A⁴ and since regression analysis has been successfully applied to structure-activity problems, we felt that F/A could also be valuable in the solution of these same problems.

In this paper, the technique of F/A will be used to examine the structure-activity relationships of a series of 21 diphenylaminopropanols on 11 biological tests, using the experimental data of Keasling and Moffett (1971).¹⁴ An attempt will also be made to analyze the relationship between the physiological host parameters assayed by different biological tests.

Review of Factor Analysis. The form of factor analysis employed here is not the standard one which appears in the statistical literature. For a discussion of the standard factor analysis methodology as used by the psychologists and statisticians, one can find several fine texts.¹⁵ These can serve the reader as an excellent introduction to this area. This standard form of factor analysis is not as efficient as regression analysis in identifying physical or chemical parameters of the data system with the causes of the observed variations in the data. Malinowski² recognized this limitation and redeveloped the formalism of factor analysis to include a form of regression analysis directly into the procedure as a factor identifier. The complete mathematical details of this modified factor analysis appear in the literature⁴ and will not be repeated here. A brief discussion will be given on the merits and drawbacks of the present approach in the solution of multidimensional problems. We would like to mention that the mathematical details need not be mastered in detail since a computer program has been written to perform all the following steps automatically. Copies of the program are available upon request. Therefore, a conceptual understanding of the following discussion should be sufficient to acquaint the reader with our approach.

If a given data set meets the necessary constraints of F/A (*i.e.*, see eq 1), and if the experimental accuracy of the data is known, then F/A can determine the number of independent parameters or factors needed to account for the observed variations in the data. At this stage of the analysis, the identity of the factors is not known, and they will be referred to as abstract factors or eigenvectors. By the present

† Page charges assisted by the chemistry department and the Research Institute of Pharmaceutical Sciences, University of Mississippi.

approach, one can determine the number of abstract factors without identifying them. This cannot be accomplished using the standard form of regression analysis as described by Hansch.^{1a} In the regression analysis approach one must identify all the factors with physically significant parameters in order to obtain an estimate of the number of factors necessary to account for the observed variations in the data. One can alternately state this by saying that factor analysis allows one to find the number of abstract factors needed to span the factor space. This is mathematically accomplished by trying to reproduce the original data array first by using only the largest abstract factor and then by systematically adding successive smaller abstract factors until the data are reproduced within experimental error. The number of abstract factors or eigenvectors needed to accomplish this is equivalent to the dimensionality of the factor space.

At this stage of the analysis, one only knows the number of factors required to reproduce the data but not their identity. With the present F/A scheme, one can identify each of the abstract factors with properties of the system (*i.e.*, drug lipid solubility, molecular weight, or polarizability). This is accomplished by setting up test vectors for each of the suspected physical or chemical properties of the system and seeing if they can be associated with the abstract factors defined previously. These test factors are exactly analogous to the test parameters used in a normal regression analysis. However, a different criterion of fit is used. A least-squares rotation matrix is generated which will attempt to mathematically rotate every point on a given abstract factor onto the suspected test factor. After the rotation has been made, each rotated value of the abstract factor should correspond to an individual value on the test factor, if the two vectors only differ by rotation in space. Therefore, one simply "reads off" the rotated values and compares each one against its suspected value. If the rotated values agree well with those of the suspected test factor, then the suspected test parameter has been identified with one of the abstract factors in the space. (This testing procedure is unique to the factor analysis model developed by Malinowski.²) One powerful advantage of the present technique is that not all points must be defined on the test factor in order to test it by the least-squares rotation scheme.⁴ A value will be predicted for all points on the abstract factor even if some of the corresponding points are missing of the test factor.

The present procedure has the drawback that one does not know which abstract factor is being rotated onto a test factor. It might be a dominant or a secondary factor. Thus, if only one test factor is identified, then one cannot determine its significance in the observed experimental data. Only if all of the abstract factors are identified with physically significant parameters is it then possible to determine a relative ordering of the importance of the individual factors.

To determine whether one has truly identified all the important factors in the factor space, an attempt is made to recalculate the original data matrix using all the suspected test factors rather than the abstract factors. Only if one has correctly accounted for all of the abstract factors will this last step be successful. Therefore, if the original data matrix can be recalculated within experimental error using all the suspected test factors, then it is almost certain that a correct solution to the problem has been found.

Finally, if both the number of rows and columns of the data matrix are greater than the number of parameters in the data set, then the same number of factors should be sufficient to span the set irrespective of whether drugs or bio-

logical activities occupy the rows of the matrix. Factor analysis is set up to test properties of the rows of the data matrix. Having the capability of testing factors associated with either the drugs or the host by simply transposing the data matrix aids greatly in the total solution of a problem.

Application of F/A to the Data of Keasling and Moffett.¹⁴ When F/A was applied to the natural logarithm of the activity data of Keasling and Moffett,¹⁴ using all the compounds in their Table II except those which lacked complete data, and omitting the dog anorexic test due to incomplete data, we found that eight abstract eigenvectors or factors are required to account for the variance in the data. The decision that eight factors are sufficient to span the data space was made by comparing the difference between the experimental data matrix and the recalculated data matrix as the number of abstract eigenvectors in the space was systematically increased. With eight factors, the experimental data are reproduced within ± 0.1 ln units. Nevertheless, the choice of eight factors is somewhat arbitrary since we do not have a reliable estimate of the standard error in Keasling and Moffett's data and since the end point of some of the biological tests is based on a subjective rating (*e.g.*, tremor and anticholinergic activities).

Before attempting to identify physical and chemical properties of the compounds with the abstract eigenvectors of the space, it is useful to perform a uniqueness test⁸⁻¹⁰ for all compounds and also for all biological activities. This test serves to characterize compounds or biological activities which possess a unique factor not present in the other compounds or activities. It can also be used to single out poor data when no rational explanation is available to explain the specific uniqueness. Mathematically, the meaning of the uniqueness test can be understood by studying the following simple equations. It is assumed that the data in all the rows of the data set except one can be accounted for by equations of the following form

$$\log \text{BA}(i, \alpha) = D(i, 1)H(1, \alpha) + D(i, 2)H(2, \alpha) + \dots + D(i, m)V(m, \alpha) \quad (3)$$


where m is one less than the number of abstract factors in the space. One assumes that the unique data row is given by the following equation

$$\log \text{BA}(U, \alpha) = D(U, 1)H(1, \alpha) + D(U, 2)H(2, \alpha) + \dots + D(U, m)H(m, \alpha) + D(U, m + 1)H(m + 1, \alpha) \quad (4)$$

where the $m + 1$ term corresponds to the unique factor. In a normal factor analysis problem, if one suspects the identity of the first factor in eq 3, then one can create a test factor consisting of either the $D(i, 1)$ or $H(1, \alpha)$ values for each case (variable) in the set. For example, the $D(i, 1)$ terms might refer to the lipid solubility of each drug, i . Similarly, to test for uniqueness for a given case (variable), a test factor is created in which all the $D(i, m + 1)$ or $H(m + 1, \alpha)$ terms are set equal to zero (see eq 3), and the $D(U, m + 1)$ or $H(m + 1, \alpha)$ terms are set equal to unity (see eq 4). The assigned unity value is arbitrary; any constant could have been used. This latter test factor will only be successfully rotated onto an abstract factor of the space if there is something truly unique about the row (column) assigned unity value on the test factor. This uniqueness test is made for each row (column) in the data set.

Another interpretation can be given to the meaning of the uniqueness test results. It is possible that all cases (variables) are given by equations in the form of eq 3, *i.e.*, no unique factor present in data. In this case, m would correspond to the number of abstract factors in the space. Now for any

Table I. Uniqueness of Test Drugs

AN  structure ^a	Value predicted ^b	Other drugs with high uniqueness relative to test drug
1 CH ₂ CH ₂ N(CH ₂) ₃ CH ₂	0.54	5 (0.21), 3, 6, 9, 14, 15 (0.10-0.15)
2 C(=CH ₂)CH ₂ N[CH(CH ₃) ₂] ₂	0.72	3 (0.23), 7, 10, 15 (0.12-0.14)
3 CH(CH ₃)CH ₂ N[CH(CH ₃) ₂] ₂	0.65	2 (0.23), 9 (0.19), 1 (0.15)
4 C(C=CH ₂)CH ₂ N(CH ₂) ₃ CH ₂	0.54	12 (0.35), 7, 14 (0.12-0.16)
5 CH(CH ₂ CH ₃)CH ₂ N(CH ₂) ₃ CH ₂	0.35	15 (0.25), 1 (0.21), 12 (0.20)
6 CH[CH(CH ₃) ₂]CH ₂ N(CH ₂) ₃ CH ₂	0.81	10 (0.23), 1 (0.14)
7 CH(C ₆ H ₅)CH ₂ N(CH ₂) ₃ CH ₂	0.43	16 (0.33), 2, 4, 9 (0.13-0.14)
8 CH[CH(CH ₃) ₂]CH ₂ NH ₂	0.33	11 (0.39), 14 (0.15)
9 C(CH ₃) ₂ CH ₂ N(CH ₂) ₃ CH ₂	0.55	14 (0.25), 3 (0.19), 1, 7, 13 (0.10-0.13)
10 CH ₂ CH(CH ₃)N(CH ₂) ₃ CH ₂	0.42	6 (0.23), 2 (0.17), 13, 14, 15 (0.12-0.14)
11 CH ₂ CH(CH ₃)NHCH ₃	0.55	8 (0.39), 14 (0.15)
12 CH(CH ₂) ₃ N(CH ₂) ₃ CH ₂	0.45	4 (0.35), 5 (0.20)
13 CH ₂ CH(CH ₃)CH ₂ N(CH ₂) ₃ CH ₂	0.35	14 (0.19), 15 (0.20), 10, 11, 5 (0.10-0.14)
14 CH ₂ CH(CH ₃)N(CH ₃)CH ₂ C ₆ H ₅	0.34	9 (0.26), 13 (0.19), 10, 11, 1 (0.10-0.15)
15 CHCH ₂ N(CH ₃)CH ₂ CH ₂	0.33	5 (0.25), 13 (0.20), 16, 2, 1 (0.12-0.14)
16 CHCH ₂ NCH ₂ CH ₂ CHCH ₂ CH ₂	0.64	7 (0.33), 15 (0.14)

^aStructural variations in the connecting link of (C₆H₅)₂C(OH)ANR₂·HX, Table II, Keasling and Moffett, p 1110, ref 14. ^bThe individual drug tested for uniqueness was assigned a value of 1.0 and the other 15 drugs were assigned a value of 0.0. The uniqueness value above is the number predicted for the test drug by this method.

given factor (such as lipid solubility), some cases (variables) have high values and others have low values. A high predicted uniqueness in this situation would mean that the row (column) in question contains a high value for the factor being tested, *i.e.*, high lipid solubility. However, since other rows (*i.e.*, drugs) may also have large values for this factor, the fit of the rotated factor will be poorer, in that several other rotated values might be relatively high even though they were assigned a zero value on the test factor in question (remember, the rotation is a least-squares one). One can sometimes differentiate between the two types of uniqueness by observing whether all the other points on the uniqueness test factor come out near zero while the row being tested for uniqueness comes out near unity. If there are no other "high" predicted values on a given uniqueness test, and there is nothing unique about the row in question (from the investigators' knowledge of their system), then one might suspect that the uniqueness may be caused by some error in the data for that row of the data set. Alternately, if a given row has a high predicted uniqueness value, but several other rows on that uniqueness test also have fairly high values, then one is probably picking out a high value on a given factor. The other high predicted values indicate that these rows of data share the same high value for this given factor, *i.e.*, high lipid solubility. Despite this arbitrariness, the uniqueness test is a useful preliminary test which allows subsequent test vectors to be chosen more intelligently.

At present, the main use of presenting the results of the uniqueness test is to provide the reader with some idea of the internal self-consistency of the data set, in the situation when he is not able to redo the data to check its validity. If a particular row is found to contain some uniqueness, whether real or caused by error, the reader should be made aware of this fact. This uniqueness is usually not obvious by a cursory examination of the data, especially if the data set is large. It is also interesting to speculate on the cause of a given true uniqueness when it is found that several rows load high on a given uniqueness test factor.

When a uniqueness test was performed for all the compounds in Table II of Keasling and Moffett's paper, it was

found that compounds II-20 and II-2 had high uniqueness values with no other high predicted values on the given test vector. Since the structures of both of these compounds did not indicate any apparent "uniqueness" from the other compounds in the scheme, it is possible that the data for these two compounds might be in error. Therefore, we eliminated them from the subsequent tests. For compound II-2 it was noted that the ED₅₀ values on tests 3, 5, and 10 were indicated as >25 mg/kg and the ED₅₀ value on test 9 was indicated as >20 mg/kg in Table I, whereas the same ED₅₀ values were indicated as simply 25 or 20 mg/kg, respectively, in Table II[†] of Keasling and Moffett's paper. Since we used the values in Table II for F/A, the high uniqueness for compound II-2 is very probably due to the large error introduced by >ED₅₀ values. We do not know if the ED₅₀ values for compound II-20 in their Table II are also in error. It is also possible that the high uniqueness of this compound is due to an unrecognized physiochemical factor. Since we could not readily determine the cause of the uniqueness for compound II-20 and since we preferred to work with a smaller factor space, we also eliminated it from the data set in this initial study.

The results of the uniqueness tests performed on the other 16 compounds are indicated in Table I. Each value for the compound of interest corresponds to a separate uniqueness test vector in which the value of that compound was set equal to one, and the values of all the other compounds were set equal to zero. To the right of the table are the other compounds which had high predicted values, when assigned zeros relative to the compound tested for uniqueness. As an example of how this table can prove useful, it can be seen that compounds 2 and 3 both had relatively high uniqueness values, 0.72 and 0.65, respectively, and that both had the high value of 0.23 relative to each other. Since these two compounds are identical structurally except for the addition of a methylene group on the β carbon in the case of compound 2 and a methyl group at that position in the case of compound 3, and since compound 3 is more potent

[†] > indicates lack of activity at the listed dose which was the highest dose tested.¹⁴

than compound 2 on 8 of the 11 biological activities, this result may indicate that the group on the β carbon influences biological activity.

Compound 6 also exhibits a high uniqueness, with compounds 1 and 10 showing high values relative to it. Since all three compounds possess a pyrrolidine ring at the nitrogen end of the molecule, this structural feature may confer a unique factor to these compounds. This feature will be tested later in trying to identify the abstract eigenvectors of the data space with chemical properties of the drugs. The low uniqueness values of the other compounds, such as 5, 8, 13, 14, and 15, indicate that these compounds are not "unique" and that they share many diverse properties with each other.

One useful feature inherent in the F/A program is the ability to transpose the rows and the columns of the data matrix before applying factor analysis and thereby test properties associated with the columns, in this case, the biological tests. Using the transposed data matrix, *i.e.*, rows corresponding to biological activities and columns to drugs, a uniqueness test was performed for each of the biological activities. The results of these tests are shown in Table II. Several interesting relationships are indicated by the data.

1. Test 1, lethality, has a high uniqueness value of 0.94, with relatedness to test 9, anticholinergic activity. From the relatedness or commonality of these two tests, it is tempting to suggest that lethality may be due to excess anticholinergic activity, centrally or peripherally.

2. The next group of tests, 2, 3, and 4, measures drug antagonism of three behavioral reflexes. It is interesting that F/A predicts commonality between two of the behavioral reflex activities, traction (test 2) and chimney (test 3), but not between reflex activity measured by the dish reflex (test 4). Since test 4 has a relatively high uniqueness value (0.89) with commonality only to test 10, it appears that this test is more closely related to anticholinergic activity, measured by pupil diameter, than to reflex activity. It also appears likely that a cholinergic component is involved in the other two behavioral reflex activities since tests 2 and 3 both show relatedness to peripheral cholinergic activity tests 9 and 10. This is not difficult to rationalize since reflex activity involves acetylcholine release at nicotinic sites and peripheral cholinergic activity involves acetylcholine release at muscarinic sites. In addition, tests 2 and 3 show relatedness to test 6, electroshock. This suggests that there is a common host parameter associated with tests 2, 3, and 6.

Table II. Uniqueness of Biological Activities^a

Test	Activity	Value predicted ^b	Other high values
1	Lethal	0.94	9 (0.11)
2	Traction	0.73	6 (0.31), 9 (0.14), 3 (0.13)
3	Chimney	0.50	10 (0.31), 6 (0.24), 2 (0.13)
4	Dish	0.89	10 (0.20)
5	TSC	0.98	6 (0.12)
6	Electroshock	0.28	2 (0.30), 3 (0.24), 5 (0.12), 10 (0.11)
7	Nicotine	0.99	
8	Tremor	0.55	9 (0.33), 11 (0.24), 10 (0.19)
9	Antichol	0.69	8 (0.33), 2 (0.14), 1 (0.11)
10	Pupil	0.57	3 (0.30), 4 (0.20), 8 (0.19), 6 (0.11)
11	Fighting	0.86	8 (0.24)

^aReference 14. ^bThe uniqueness value predicted above is the value predicted when the single biological activity tested is assigned a value of 1.0 and the other 10 biological activities are assigned the value of zero.

Table III. Prediction of the Relatedness of Anticonvulsant Tests^a

Test	Activity	Test	Predicted	Test	Predicted ^b
1	Lethal	0	0	0	0
2	Traction	0	0	0	0
3	Chimney	0	0	0	0
4	Dish	0	0	0	0
5	TSC	1.00	0.994	---	0.314
6	Electroshock	---	0.202	---	0.064
7	Nicotine	---	3.065	1.0	0.994
8	Tremor	0	0	0	0
9	Antichol	0	0	0	0
10	Pupil	0	0	0	0
11	Fighting	0	0	0	0

^aProcedures: thiosemicarbazide antagonism, electroshock antagonism, and nicotine-seizure antagonism. ^b0 for predicted values indicates values less than 0.01.

3. Tests 5, 6, and 7 are used to measure anticonvulsant activity against convulsions or death induced by thiosemicarbazide, electroshock, and nicotine, respectively. Of the three tests, test 7 showed very high uniqueness with no relatedness to other tests. As mentioned before, a high uniqueness value without a rational explanation may point to erroneous data. This appears to be unlikely (R. B. Moffett, personal communication), so the high uniqueness of nicotine may be related to the mechanism by which nicotine induces convulsions and the manner in which the drugs in our data set antagonize these convulsions. On the other hand, test 6 has a very low uniqueness value, with relatedness to diverse tests in all categories of biological activity. This low uniqueness value indicates that antagonism of electroshock convulsions may not be a selective anticonvulsant screening test, as it contains host parameters related to cholinergic activity (test 10) and to reflex activity (tests 2 and 3). Thus, F/A has shown us mathematically that this test is measuring several host parameters. Test 5 with a high uniqueness value of 0.98 and with relatedness only to test 6 appears to be a more specific anticonvulsant test. However, our results indicate that anticonvulsant activity measured against nicotine-induced convulsions is different from convulsions induced by thiosemicarbazide or by electroshock.

4. Tests 8, 9, and 10 grouped together as anticholinergic tests by Keasling and Moffett¹⁴ showed broad relatedness to the other tests and did not exhibit high uniqueness values. F/A has shown us a new method of looking at biological parameters. From this approach, we can gain a better understanding of the interrelatedness of various biological screening procedures and possibly gain insight into their usefulness in detecting new drugs.

The relative relatedness of biological activities can be carried one step further. A crude rating of relatedness can be obtained by assigning a value of unity to the specific biological activity with a high uniqueness value, assigning zeros to those biological tests lacking the unique host parameter, and not assigning numbers to the tests which showed relatedness to the unique test. The F/A program will then predict values for these unassigned or free-floating points relative to the 0 and 1.0 scale. In setting up a ranking scheme it is important to assign enough true zeros so as not to bias the test results. For biological tests which share a broad commonality with many other tests, the test vector cannot be easily defined. Therefore, the only group of related biological tests that could be used in ranking schemes of this sort is the anticonvulsant tests 5, 6, and 7. In an eight-factor problem, F/A requires that at least nine points are defined numerically on the vector; therefore, only 2 tests can be free-floating. The results of ranking of the

Table IV. Identification of Abstract Eigenvectors with Structural Parameters of Test Drugs

Drug	Ring on N ^a	
	Test	Predicted
1	1.0	1.08
2	0	0.02
3	0	0.001
4	1.0	0.99
5	1.0	0.91
6	1.0	0.98
7	1.0	1.05
8	0	0.01
9	1.0	0.93
10	---	0.46
11	0	0.01
12	---	1.00
13	1.0	1.00
14	---	0.09
15	---	0.69
16	---	1.16

^a1.0 indicates the presence of a ring at the nitrogen end of the molecule, 0 indicates the absence of a ring at the nitrogen end of the molecule, and --- indicates that no value was assigned on the test vector.

relatedness of tests 5, 6, and 7 are shown in Table III. In the first column, test 5 was assigned a value of 1.0 and tests 6 and 7 were free-floated. F/A predicted the values of 0.202 and 3.065 for tests 6 and 7, respectively. These results indicate that the common host parameter (in this case, possibly the active sites needed for induction of convulsions) which is being affected by these three tests is more active in test 7 than in tests 5 and 6, with the order $7 > 5 > 6$. Basically the same results were obtained when test 7 was assigned the value of 1.0 and tests 5 and 6 were free-floated.

Testing Structural Parameters for Possible Identification with the Abstract Eigenvectors. Once the number of controlling factors has been determined, it is possible to test various structural parameters of the drugs for possible identification with the abstract eigenvectors. Using the data matrix with the compounds as rows, several structural features of the molecules were tested as possible factors. A specific structural parameter can be tested crudely by assigning values of unity to compounds possessing the parameter and assigning values of zero to all compounds lacking the structural feature, or, if available, more quantitative values can be assigned to the compounds. The former method is similar to that originally proposed by Free and Wilson.¹⁶ F/A will then analyze the test vector and predict values for all points on the vector. If the structural parameter under question is a true factor, the predicted values will agree closely with the assigned values. Results from the preliminary uniqueness tests (Table I) suggest that the presence or absence of the pyrrolidine ring on the molecule might test well as a factor. Therefore, we assigned the value of unity to compounds possessing the pyrrolidine ring and the value of zero to those lacking it. We also allowed some compounds to free-float. As can be seen from the results in Table IV, the agreement of the predicted values with the assigned values is quite good. In addition, the values predicted for compounds which were free-floated agree with the structural features of the molecules. Thus, for compound 12 possessing the ring, there was a predicted value of 1.00, and for compound 14, lacking the ring, there was a predicted value of 0.09. Compounds 15 and 16 both contain complex rings different from the pyrrolidine ring. It is interesting that F/A predicts the values of 0.69 and 1.16 for these two compounds, respectively. These values indicate

that compounds 15 and 16 share a similar structural parameter with the pyrrolidine ring. For compound 10, F/A predicted the value of 0.46 although the compound has a pyrrolidine ring. This rather low value may indicate that the presence of the ring in compound 10 is not equivalent to its presence in the other compounds. Since compound 10 is the only pyrrolidine ring-containing compound with a methyl group attached to the carbon adjacent to the nitrogen of the ring, the influence of the ring on biological activities may be altered. In any case, with the minor exception of compound 10, the presence of the pyrrolidine ring tested well as a factor.

The ability to test a single structural modification as a factor is one of the most useful features of our F/A program. In a problem where several modifications are made simultaneously to a series of drugs, it is still possible to test whether a single structural feature is exerting an important influence on the biological activities. This testing ability is quite useful in analyzing large data sets where the observed activity is a sum over many diverse effects, and it is not always obvious which structural features of the drugs are important by a simple observation of the data. For large data sets, the visual searches for specific effects can become quite difficult.

Structural features of the drugs which did not test well as factors include the following: (1) presence or absence of an isopropyl group in the molecule, (2) number of carbons between the diphenyl and amino ends of the molecule, and (3) the presence of a group on the carbon adjacent to the diphenyl end of the molecule. However, the presence or absence of a group on the carbon adjacent to the nitrogen end of the molecule tested fairly well as a factor. Although qualitative these results all indicate that the biological activity of the drugs tested resides at the nitrogen end of the molecule. The synthesis of new structural analogs could be aided by this knowledge. In a problem in which the accuracy of the data is known, more quantitative drug parameters, such as lipid solubility or electron density, could be tested as factors. Unfortunately, data of known accuracy have only been simultaneously evaluated on only one or two biological activities. Therefore, one will have to wait until a data set appears in which enough different biological activities are quantitatively evaluated before a comparison can be made between the results of regression analysis and F/A.

Factor Analysis as a Data Predicting Tool. Factor analysis can yield useful information even when all the abstract factors have not been identified with physically significant parameters.^{4,8,11} It is always possible to rotate a column of experimental data into an abstract factor of the space.^{4,7,10,12} Then, it should be possible to find eight columns of data which separately, or in conjunction, contain all eight important drug-host interaction parameters. These, then, are the "key" biological tests needed to characterize the given drugs, and all other biological activities for a given drug should be predictable as a linear combination of the key activities. Mathematically, this is given by

$$\log \text{BA}(i, \alpha) = \sum_{j=1}^N C(j, \alpha) B(i, j) \quad (5)$$

where $\text{BA}(i, \alpha)$ is the α th biological activity associated with new drug; i and $C(j, \alpha)$ are constants calculated for each biological test, α , from the F/A reproduction scheme; and $B(i, j)$ is the j th measured key biological activity of drug i . This procedure has been used successfully to predict proton chemical shifts,⁴ activity coefficients determined by gas-liquid chromatography,⁷ and solutes' retention time

Table V

Remaining tests to be predicted	Lethal	Traction	Dish	TSC	Nicotine	Tremor	Antichol	Fighting
A. Coefficients for the Drugs on the 8 Key Tests ^a								
Electroshock	0.0791	0.6665	0.1802	0.0330	0.0761	0.1723	-0.1663	-0.0564
Pupil	0.3095	0.0583	0.6357	-0.3231	0.0875	0.3175	0.0713	-0.0645
Chimney	0.2844	0.5227	0.2823	-0.3036	0.1363	0.3261	-0.2470	-0.0132
B. Coefficients for the Drugs on 7 Key Tests and the Identified Test Parameter ^a								
Pyrrolidine ring								
Electroshock	0.1120	0.7557	-0.1661	0.1622	0.0941	0.1252	-0.1653	-0.1267
Pupil	0.4255	0.3726	-0.5860	0.1326	0.1509	0.1517	0.0748	-0.3125
Chimney	0.3358	0.6623	-0.2602	-0.1013	0.1645	0.2524	-0.2454	-0.1234
Dish	0.1824	0.4947	-0.9218	0.7168	0.0998	-0.2608	0.0054	-0.3902

^aCoefficients to be substituted into eq 2 for the prediction of the data in Table VI.

Table VI. Comparison between Experimental and Predicted Log Concentrations

Drug no. ^a	AN structure	Using all 8 key test columns						Using 7 key test columns and the pyrrolidine ring test factor								
		Chimney		Electroshock		Pupil		Drug no. ^a	Chimney		Dish		Electroshock		Pupil	
		Exptl	Pred	Exptl	Pred	Exptl	Pred		Exptl	Pred	Exptl	Pred	Exptl	Pred	Exptl	Pred
1-8	-N[(CH ₂) ₃ CH ₃] ₂	4.14	4.28	5.30	4.71	5.30	4.16	1-8	4.14	4.75	3.81	5.45	5.30	5.00	5.30	5.21
1-21	-N(CH ₂) ₂ CH ₂	3.91	3.49	3.22	3.31	3.91	3.78	1-21	3.91	3.28	3.00	1.23	3.22	3.17	3.91	3.30
1-23	-N(CH ₂) ₂ CH ₂	3.91	3.58	3.33	3.25	4.27	3.48	1-23	3.91	3.27	2.64	1.32	3.33	3.05	4.27	2.77
11-6	-NH ₂	1.72	2.21	1.61	1.30	3.21	2.58	11-6	1.72	2.35	0.69	1.21	1.61	1.39	3.21	2.92
11-9	-NH ₂	3.21	2.83	3.22	2.65	3.21	2.98	11-9	3.21	2.32	1.61	3.37	3.22	2.97	3.21	4.10
11-18	-NHCH ₃	2.53	3.48	2.99	2.90	3.22	3.62	11-18	2.53	3.59	1.95	2.33	2.99	2.96	3.22	3.86
11-23	-NHCH ₂ CH ₃	2.30	3.16	2.99	2.98	3.22	3.42	11-23	2.30	3.43	2.07	3.03	2.99	3.16	3.22	4.02
11-33	-NHCHCH ₂ CH ₂	3.91	3.37	3.91	3.44	3.91	3.67	11-33	3.91	3.62	2.77	3.24	3.91	3.59	3.91	4.18
11-40	-NHCH(CH ₂) ₂ CH ₂		3.59	3.46	3.28	3.91	3.71	11-40		3.42	2.56	2.28	3.46	3.22	3.91	3.47
111-1	-N(CH ₃) ₂	3.22	3.51	3.00	3.18	3.91	3.87	111-1	3.22	3.27	2.89	2.07	3.00	3.04	3.91	3.34
111-2	-N(CH ₂ CH ₃) ₂	3.22	3.62	3.22	3.18	3.91	4.19	111-2	3.22	3.14	3.13	2.36	3.22	3.04	3.91	3.69

^aCode letter from Tables I and III, Keasling and Moffett, pp 1108, 1111, ref 14.

or index in gas-liquid chromatography.^{9,10,12} To work well, it is critical that accurate data are obtained on the key biological tests for the compound for which predictions are to be made. The results of the present section are only fair to poor. They are presented mainly to show the reader another way in which F/A may be useful in biological problems.

To find the eight key biological activities needed to best span the data space, all combinations of 8 of 11 biological activities were tried. For each set, the deviation between the original and the recalculated data was computed, the criteria for best fit being the lowest calculated average deviation. The best combination of test columns which contained all eight important parameters and which yielded an overall mean error of 0.15 log units was: (1) lethal, (2) traction, (4) dish, (5) TSC, (7) nicotine, (8) tremor, (9) anticholinergic, and (11) fighting. Using the measured biological activities of the drugs on these eight key activities simultaneously as test factors, a rotation was made, and equations in the form of eq 5 were generated which should predict the α th biological activity on the i th drug in terms of the eight key measured biological activities of the i th drug. The coefficients shown in Table VA are only those for the three activities not included among the 8 of 11 activities used as key tests. For the eight test columns, $C(j, \alpha) = 1$ for $j = \alpha$ and for $\alpha \neq j$, $C(j, \alpha) = 0$. The equations, generated using the data in Table II of Keasling and Moffett's paper, are then tested on several compounds from Tables I and III of their paper which were not included in our original data matrix. These results are shown in Table VI. The results for electroshock and pupil are good considering the accuracy of the experimental data of Keasling and Moffett. Many of their values are only reported as >50 or <10 . As mentioned above, this type of

value introduces a large error. Since values similar to these are included in both the data matrix itself and the compounds used as test subjects in Table VI, quantitative agreement is not expected.

As one identifies physically significant parameters with the abstract factors of the space, it is then legitimate to replace one of the biological test columns used as the key tests with this physically significant parameter reducing the number of biological activities needed to characterize a new drug. We have previously shown (Table IV) that the presence or absence of a pyrrolidine ring seemed to test as a drug factor. This chemically significant parameter was substituted for each of the eight key tests, and the agreement between the original data matrix and the recalculated data matrix was noted. It was found that replacing test 4 (dish) with the pyrrolidine ring test vector yielded a mean overall error of 0.13 log units. A new set of coefficients for those biological activities not included with the "best seven" key tests was generated and is shown in Table VB. The results of using these coefficients in eq 2 for the left-out activities are shown in Table VI. The set was moderately worse, as expected, but the same trends were predicted. Furthermore, from noting the sign of the coefficient corresponding to the pyrrolidine ring test, we can see that they are all negative. This indicates that the presence of the pyrrolidine ring reduces the amount of drug necessary to produce a given response.

In conclusion, it was shown that factor analysis can be effectively applied to the area of drug structure-biological activity relationships. Important information can be gained from this new approach. F/A can pinpoint erroneous data. F/A allows isolated physical or chemical properties of the

drugs to be tested to see if they are influential determinants of biological activity. Even if none of the abstract factors are identified, F/A can be useful in choosing the key columns of data which encompass all independent drug-host interactions. From these key columns of data, one can predict the other columns of data for a new drug. Another asset of F/A is the potentially valuable property of examining the interrelatedness of the biological tests. From this information we can gain a greater insight into the physiological responses and the mechanisms of drug action.

Acknowledgments. The authors wish to thank Dr. Robert Mikeal, Associate Professor of Pharmacy Administration, for his many valuable discussions. We would also like to acknowledge the help of the Computer Center of the University of Mississippi and the Pharmacy Data Center of the School of Pharmacy for access to their computer facilities. One of us (M. L. W.) wishes to thank the Research Institute of Pharmaceutical Sciences for supporting this work.

References

- (1) (a) C. Hansch, *Accounts Chem. Res.*, **2**, 232 (1969); (b) C. Hansch and W. R. Glave, *Mol. Pharmacol.*, **7**, 337 (1971); (c) A. Cammarata and K. S. Rogers, *J. Med. Chem.*, **14**, 269 (1971); (d) A. Cammarata, *ibid.*, **15**, 573 (1972).
- (2) E. R. Malinowski, Ph.D. Thesis, Stevens Institute of Technology, Hoboken, N. J., 1961; *Diss. Abstr. B*, **23** (1963).
- (3) P. H. Weiner, Ph.D. Thesis, Stevens Institute of Technology, Hoboken, N. J., 1971; *Diss. Abstr. B*, **32** (1971).
- (4) P. H. Weiner, E. R. Malinowski, and A. R. Levinstone, *J. Phys. Chem.*, **74**, 4537 (1970).
- (5) P. H. Weiner and E. R. Malinowski, *ibid.*, **75**, 1207 (1971).
- (6) P. H. Weiner and E. R. Malinowski, *ibid.*, **75**, 3160 (1971).
- (7) P. T. Funke, E. R. Malinowski, D. E. Martire, and L. Z. Pollara, *Separ. Sci.*, **1**, 661 (1967).
- (8) P. H. Weiner and D. G. Howery, *Can. J. Chem.*, **50**, 448 (1972).
- (9) P. H. Weiner and D. G. Howery, *Anal. Chem.*, **44**, 1189 (1972).
- (10) P. H. Weiner, C. Dack, and D. G. Howery, *J. Chromatogr.*, **69**, 249 (1972).
- (11) P. H. Weiner and J. F. Parcher, *J. Chromatogr. Sci.*, **10**, 612 (1972).
- (12) P. H. Weiner and J. F. Parcher, *Anal. Chem.*, **45**, 302 (1973).
- (13) A. Gautier, J. Zurli, R. C. Cros, and H. Sarles, *Eur. J. Clin. Biol. Res.*, **17**, 574 (1972).
- (14) H. H. Keasling and R. B. Moffett, *J. Med. Chem.*, **14**, 1106 (1971).
- (15) (a) R. J. Rummel, "Applied Factor Analysis," Northwestern University Press, Evanston, Ill., 1970; (b) R. B. Catell, "Factor Analysis," Harper and Row, New York, N. Y., 1952; (c) D. N. Lawley and A. E. Maxwell, "Factor Analyses as a Statistical Method," Butterworths, London, 1963.
- (16) S. M. Free and J. W. Wilson, *J. Med. Chem.*, **7**, 395 (1964).

Structure-Activity Correlations of Antimalarial Compounds. 2. Phenanthreneaminoalkylcarbinol Antimalarials

Paul N. Craig*

Craig Chemical Consulting Services, Inc., Ambler, Pennsylvania 19002

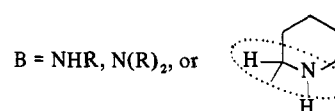
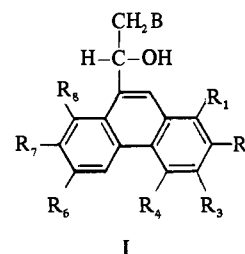
and Corwin H. Hansch

Department of Chemistry, Pomona College, Claremont, California 91711. Received November 27, 1972

The antimalarial structure-activity relationships in a series of phenanthreneaminoalkylcarbinols have been studied by both the additivity (or Free-Wilson) and multiple parameter analysis methods. Both methods agree on a major finding: whereas the 1-octanol-water partition constants of substituents in the aromatic rings (position 1-8) correlate well with the antimalarial data, the far greater variations in partition coefficients of the aminoalkyl groups do not correlate at all with the biological data. This finding results from a multiple parameter analysis with 54 of 60 analogs for which data were available in 1970 and from additivity analyses with 43 and 28 analogs. In 1971, 47 more analogs were tested, and from a study of 102 of 107 analogs, a separation of polar and partition effects was possible. To obtain these latter results, a redetermination of the partition coefficient for 4-trifluoromethylphenoxyacetic acid was made, and the revised π value for the aromatic CF_3 group is 0.88 log P units.

In support of the increasing emphasis on the development of more effective antimalarial agents, the application of computerized regression analysis to a study of chemical structure-antimalarial activity relationships was begun in 1969 under contract with the Walter Reed Army Institute of Research.¹

A. Multiple Parameter Analyses. Antimalarial test results for 60 phenanthreneaminoalkylcarbinols of structure I were examined in Sept 1970 by the multiple parameter method of analysis.² After converting the animal experimental data³ to estimated ED_{50} values, log $1/C$ values were calculated, where C is the concentration of test drug in moles per kilogram of test animal. These quantitative expressions of antimalarial activity were examined by regression analyses for correlations with various combinations of the following parameters: π_{sum} , π_x , π_y , π_{x+y} , σ_x , σ_y , σ_{x+y} (see Table I footnotes for definitions). In estimating π_{sum} values, the amino side chain was treated as follows. The $-\text{CHOHCH}_2\text{N}$ moiety was assumed to be constant, and the π values for R or R_2 were used. For the 2-piperidyl group,



the encircled moiety was considered to be equivalent to the CH_2N group, and the π value for four cyclohexane methylene groups (4×0.42) was used. From this was subtracted 0.13 for the branching at the 2 position of the piperidine ring; thus 1.55 was used for the π contribution due to the 2-piperidyl group, 4.0 and 7.0 for N -butyl₂ and N -heptyl₂, respectively.⁴