enstaff, in "Abstracts of the 173rd ACS National Meeting," New Orleans, La., Mar. 1977, ORGN 13.

(2) H. Seyle, I. Meces, and S. Szabo, *Int. Urol. Nephrol.*, **2,** 287 (1970). H. Seyle, *Science*, **169,** 775 (1970).

(3) V. Eyble, J. Sýkora, M. Koutenská, J. Koutenský, and F. Mertl, *Arzneim.-Forsch.*, **23,** 867 (1973).

(4) J. Strating and H. J. Backer, *Rec. Trav. Chim.*, **69,** 909 (1950).

(5) G. L. O'Connor and H. R. Nace, *J. Am. Chem. Soc.*, **75,** 2118 (1953).

(6) F. C. Chang, R. T. Blickenstaff, A. Feldstein, J. R. Gray, G. S. McCaleb, and D. H. Sprunt, *ibid.*, **79,** 2164 (1957).

(7) L. C. King, R. M. Dodson, and L. A. Subluskey, *ibid.*, **70,** 1176 (1948).

(8) A. Bowers, L. C. Ibáñez, E. Denot, and R. Becerra, *ibid.*, **82,** 4001 (1960).

(9) D. A. Lightner and C. Djerassi, *Steroids*, **2,** 583 (1963).

(10) J. E. Herz and J. Fried, U.S. pat. 2,842,568 (July 9, 1958); through *Chem. Abstr.*, **53,** 456i (1959).

(11) D. A. Swann and J. H. Turnbull, *Tetrahedron*, **20,** 1265 (1964).

(12) R. T. Blickenstaff and F. C. Chang, *J. Am. Chem. Soc.*, **80,** 2726 (1958).

(13) C. R. Engel and G. Just, *ibid.*, **76,** 4909 (1954).

(14) B. Bannister and F. Kagan, *ibid.*, **82,** 3363 (1960).

(15) J. H. Turnbull, *Chem. Ind.*, **1959,** 515.

(16) J. Kawanami, Japanese pat. 8348 (1966); through *Chem. Abstr.*, **65,** 10641f (1966).

(17) F. A. Kincl, *Chem. Ber.*, **93,** 1043 (1960).

(18) A. Segaloff and R. B. Gabbard, *Steroids*, **5,** 219 (1965).

(19) D. A. Swann and J. H. Turnbull, *Tetrahedron*, **22,** 231 (1966).

(20) A. K. Bahn, "Basic Medical Statistics," Grune & Stratton, New York, N.Y., 1972, pp. 52–55.

(21) "Physicians' Desk Reference," 22nd ed., Medical Economics Inc., Oradell, N.J., 1968, p. 1063.

## ACKNOWLEDGMENTS

Supported by PHS-R01-ES-942-03 and by Veterans Administration Research Project 8-583.

The spironolactone was a gift of G. D. Searle and Co.

# Molecular Connectivity and Substructure Analysis

**LOWELL H. HALL** *x and **LEMONT B. KIER** ‡

Received February 21, 1978, from the **Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170,* and the *‡Department of Pharmaceutical Chemistry, Medical College of Virginia, Richmond, VA 23298.* Accepted for publication May 19, 1978.

**Abstract** □ Antimicrobial and antiviral data sets were analyzed by molecular connectivity. Standard structure–activity relationship equations of high quality were produced in both cases. For phenyl propyl ether activity against *Staphylococcus aureus*, the three variables ${}^1\chi$, ${}^3\chi_P$, and ${}^4\chi^v_{PC}$ yielded an *r* of 0.957, significantly better than a $\pi,\sigma$ analysis. Analysis of benzimidazole antiviral data (Lee strain, B flu virus) revealed that the one variable, ${}^6\chi_P$, yielded an *r* of 0.950, also better than a reported Hansch analysis. Both data sets were further analyzed by partitioning the important regression variables into terms representing various structural features of the molecules. For the phenyl propyl ethers, the *para*-region of the phenyl ring is important for improved activity and the negative coefficient on ${}^3\chi_P$ corresponds to decreased activity for *vic*-dihydroxy compounds. For the alkylbenzimidazoles, substitution on the five-membered ring is highly important. No discrimination of six-membered ring positions was revealed. These structure–activity relationship observations can form the basis for synthetic decisions to improve activity.

**Keyphrases** □ Molecular connectivity—antimicrobial and antiviral data sets analyzed, structure–activity relationships produced □ Structure–activity relationships—antimicrobial and antiviral data sets analyzed by molecular connectivity method □ Antimicrobial data sets—analyzed by molecular connectivity method, structure–activity relationships produced □ Antiviral data sets—analyzed by molecular connectivity method, structure–activity relationships produced □ Topological indexes—molecular connectivity, antimicrobial and antiviral data sets analyzed, structure–activity relationships produced

Before there was any serious effort to quantify structure–activity relationships, it was popular to draw structural conclusions in terms of molecular fragments. Thus, combinations of atoms and groups were judged essential, based on studies of series of molecules with similar biological actions (1–3). This approach attempted to convey direct structural information to permit the familiar iterative process: synthesis → test → synthesis → test.

## BACKGROUND

These fragment proposals are certainly in an understandable and convenient form. However, the results of structure–activity relationship analyses frequently are stated in terms of values of physical properties, information far less valuable to the synthetic chemist. This approach has led to a greater quantification of structure–activity relationship results but also a loss of more useful information, *i.e.*, information directly related to structures.

The basic problem has been the lack of a numerical structural description that can be applied universally to all fragments and that can be translated from numbers back to structural fragments. Such a method, molecular connectivity (4–11), is now available for the numerical representation of molecular structure in a form suitable for multiple regression analysis. The molecular connectivity indexes give quantitative expression to structural variations described traditionally in such qualitative terms as branching, cyclization, and bond type, as well as number and kind of atoms. Structural information is encoded in the set of connectivity indexes, which may be calculated for any molecular structure.

The molecular connectivity indexes, referred to as $\chi$ indexes, are based on the hydrogen-suppressed graph; this graph is simply the molecular skeleton, including all atoms (except hydrogen) and the bonds between them. Information describing the molecular structure is extracted in numerical form from the connectivity relationships in the hydrogen-suppressed graph (4).

This paper discusses methods of identifying molecular fragments that the structure–activity relationship regression equation suggests are important. In a systematic fashion, the important molecular fragments or atomic arrangements may be identified. In this further development of the connectivity method, a high quality regression equation first is established with one or more $\chi$ indexes. Each $\chi$ index is a summation of subgraphs of a specific type. Then the subgraphs may be divided into chemically sensible sets. By regression analysis, the most important sets are picked out. Based on the topology of the important subgraphs, structure–activity relationship conclusions may be drawn as an aid to drug design.

**Formalism of Method**—Each ${}^m\chi_t$ is a sum of terms called subgraph terms:

$$ {}^m\chi_t = \sum {}^mS_j \qquad \text{(Eq. 1)} $$

Each term ${}^mS_j$ is defined for a subgraph—a skeletal fragment of *m* bonds arranged in a particular fashion (type *t*)—and the summation is over all such subgraphs in the hydrogen-suppressed graph. The symbol *m* is called the order or the index of the subgraph. The zero-order subgraph is simply a skeletal atom (graph vertex), and ${}^0\chi$ is a sum of the numerical values computed for each atom. First-order subgraphs are the skeletal bonds, and $^1\chi$ is a sum of values computed for each bond.

Second-order subgraphs consist of pairs of adjacent bonds, and $^2\chi$ is a summation of the values for all distinct sets of adjacent pairs of bonds in the skeleton. The number of second-order subgraphs (paths of length two) in a graph depends heavily on the complexity of skeletal branching. For example, in normal alkanes, the number of second-order subgraphs is $n - 2$. The count of second-order paths increases steadily with branching complexity, so that in 2,2,3-trimethylbutane there are nine compared to five in *n*-heptane.

There are three types of third-order subgraphs: path ($t = P$), cluster ($t = C$), and chain ($t = CH$) (see Ref. 4, chap. 3). Third-order path subgraphs consist of three consecutive skeletal bonds (*n*-butane skeleton). Third-order cluster subgraphs consist of three skeletal bonds sharing a common skeletal atom as in the isobutane skeleton. And third-order chain subgraphs consist of three bonds joined as a triangle (cyclopropane skeleton).

For fourth order, a new type of subgraph is encountered in addition to path (*n*-pentane skeleton), cluster (neopentane skeleton), and chain (cyclobutane skeleton). The new type is path–cluster ($t = PC$) and corresponds to the isopentane skeleton. The number of such types depends heavily on the number of three-way and four-way branch points, as well as the adjacency of such branch points. Higher order $\chi$ indexes, $^m\chi_t$, may be defined analogously.

As this brief discussion emphasizes, the concept of molecular connectivity is based on a point of view that begins with the molecular skeletal structure and develops the structural information in systematic numerical manner. A central aspect is the recognition and identification of fragments as subgraphs.

The numerical value computed for each subgraph depends on the vertex delta value, $\delta$, the number of bonded neighbor atoms, for each skeletal atom:

$$^mS_j = \prod_{i=1}^{m=1} \delta_i^{-1/2} \qquad \text{(Eq. 2)}$$

The subgraph term is the reciprocal square root of the product of the vertex delta values, $\delta_i$, for all atoms in the subgraph. For example, for the zero- and first-order indexes:

$$^0\chi = \sum_{i=1}^{N_a} \delta_i^{-1/2} \qquad \text{(Eq. 3)}$$

and:

$$^1\chi = \sum_{k=1}^{N_e} (\delta_i\delta_j)_k^{-1/2} \qquad \text{(Eq. 4)}$$

where $N_a$ is the number of atoms and $N_e$ is the number of edges in the whole graph and $k$ refers to the $k$th bond.

Identification and recognition of subgraphs depend only on the connections (bonds) in the molecular skeleton, not the identity of the individual atoms. For example, the isobutane skeleton, as well as those of trimethylamine and isopropyl alcohol, correspond to a cluster-3 subgraph. The simple connectivity index, $^3\chi_c$, is independent of the atom types and has a value of 0.577 for the three molecules. However, there is also a $\chi$ cluster index defined as above but depending on the specific identity of the atoms in the skeleton. The atom type is entered into the calculation through a modification of the vertex delta value. Atoms such as nitrogen, oxygen, and fluorine influence properties through the number of valence electrons and attached hydrogens. Thus, the valence delta is defined as:

$$\delta^v = Z^v - h \qquad \text{(Eq. 5)}$$

being the difference between the number of valence electrons and the number of hydrogen atoms bonded to the atom[1]. Thus, $\delta^v$ is a count of valence electrons involved in lone pairs and skeletal bonding. Chi indexes based on valence delta values are symbolized with the superscript $v$ as in $^0\chi^v$, $^1\chi^v$, $^4\chi^v_{PC}$, *etc.* For atoms beyond the first row, the inner shell electrons are included as follows:

$$\delta^v = (Z^v - h/Z - Z^v) \qquad \text{(Eq. 6)}$$

Each $\chi$ index is a weighted count of the specific type of subgraph (skeletal fragment). The summation includes a contribution from each distinct subgraph in the molecule. This method then replaces a separate listing of all fragments with a single numerical value by assigning a weight, $^mS_j$, to each fragment. For example, the following path four fragments may appear in a given molecule in a biological study: $HOCH_2CH_2CH_2$, $CH_3OCH_2CH_2$, $CH_3CH_2CH_2CH_2$, $ClCH_2CH_2CH_2$, $CH_3NHCH_2CH_2$, *etc.* Each fragment may be assigned a value, $^4S_P$, and summed to give the $^4\chi^v_P$ value.

In published molecular connectivity studies, the various $^m\chi_t$ indexes have been used in many high quality correlations, singly and in sets of two, three, or more. This present study shows how the $^m\chi_t$ indexes in a high quality correlation may be dissected to determine which of the contributing subgraphs are most important in the molecules under investigation. This isolation of a portion of the molecule or a particular arrangement of substituents is a systematic identification of a fragment of significance. A quantification of the fragment influence is also determined on the basis of a regression equation.

Two sets of biological data were examined, one of antimicrobial activity and the other of antiviral activity.

**Antimicrobial Activity of Phenyl Propyl Ethers**—A number of ethers of glycerol and trimethyleneglycol were prepared and tested against skin fungi (12). Chlorphenesin, 3-*p*-chlorophenoxy-1,2-propanediol, has been marketed as an antifungal agent[2], but other agents in this chemical class have considerably greater antifungal potency (12). In this study, 28 alkyl- and chloro-substituted phenyl propyl ethers were examined for their action against *Trichophyton mentagrophytes.*

The phenyl propyl ethers in this data set present an interesting problem from the standpoint of structure–activity relationships. Five positions on the phenyl ring and two on the ether side chain were modified to develop the series shown in Table I. Hansch and Lien (13) used partition coefficients and Hammett $\sigma$ terms to evaluate a property–activity relationship. Their analysis led to a modest correlation but only after the deletion of two molecules from the data set.

In the present study, the antifungal activity of these molecules was examined using molecular connectivity (4). In addition, the different skeletal patterns in the list are described by evaluating the significant contributions of specific molecular structure features as reflected in molecular connectivity indexes for salient subgraphs.

**Antiviral Activity of Alkyl-Substituted Benzimidazoles**—Sixteen alkyl-substituted derivatives of benzimidazole were examined for their effectiveness against the Lee strain of influenza B virus (Table II). Only one compound in the series studied by Tamm *et al.* (14) was an *N*-alkyl derivative. Since this compound, *N*-methylbenzimidazole, has significantly different acid–base properties and may act *via* a different mechanism, it was not considered in the present study (although its inclusion does not adversely affect the correlation statistics). Alkyl groups were substituted on all five available carbon atoms in the benzimidazole nucleus. Substituents included methyl, ethyl, propyl, isopropyl, and butyl in various combinations. The inhibitory concentrations covered over a 40-fold range.

## RESULTS

The general procedure for application of molecular connectivity was the same for both sets of biological data. Chi indexes through the sixth order were computed both for simple and valence connectivity indexes. After checking the molecular structure input data—the connectivity matrix and numbers of valence electrons—for accuracy, the following indexes were stored on computer file for further analysis: $^0\chi$, $^1\chi$, and path terms $^2\chi$ through $^6\chi_P$, $^3\chi_C$, $^4\chi_{PC}$, and $^6\chi_{CH}$, as well as the corresponding valence indexes for a total of 20 variables.

With standard multiple linear regression computer programs, all possible one- and two-variable equations were computed for both data sets. In addition, three-variable equations were examined for the phenyl propyl ether set.

**Phenyl Propyl Ether Antifungal Data Set**—Examination of the three-variable multiple regressions reveals that two equations are of very high quality and give good account of the experimental data. The two equations differ in one variable:

$$\log(1/c) = 2.44\ (\pm 0.09)^1\chi - 3.29\ (\pm 0.09)^3\chi_P + 2.71\ (\pm 0.03)^4\chi^v_{PC} - 1.31\ (\pm 2.4) \qquad \text{(Eq. 7)}$$

$$r = 0.957 \quad s = 0.149 \quad F = 87.4 \quad n = 28$$

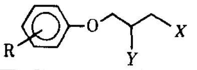
$$\log(1/c) = 1.30\ (\pm 0.10)^1\chi - 2.70\ (\pm 0.08)^3\chi_P + 2.74\ (\pm 0.04)^3\chi^v_P + 0.01\ (\pm 2.5) \qquad \text{(Eq. 8)}$$

$$r = 0.955 \quad s = 0.152 \quad F = 83.8 \quad n = 28$$

[1] For fluorine, the negative sign is taken for the square root in $^mS_j$ so that the sign of $^mS_j$ is negative for any subgraph containing fluorine.

[2] Mycil.

Table I—Antimicrobial Activity for Substituted Phenyl Propyl Ethers



| R | $X$ | $Y$ | ${}^1\chi$ | ${}^3\chi_P$ | ${}^4\chi^v_{PC}$ | Observed | Calculated[a] | Difference |
|---|---|---|---|---|---|---|---|---|
| 2-Methyl | OH | OH | 6.236 | 4.016 | 0.505 | 2.26 | 2.07 | 0.19 |
| 2-Methyl | OH | H | 5.698 | 3.404 | 0.469 | 2.46 | 2.68 | −0.22 |
| 2-Methyl | H | OH | 5.842 | 3.468 | 0.394 | 2.79 | 2.62 | 0.17 |
| 2-Chloro | OH | OH | 6.236 | 4.016 | 0.560 | 2.31 | 2.22 | 0.09 |
| 2-Chloro | OH | H | 5.698 | 3.404 | 0.524 | 2.84 | 2.83 | 0.01 |
| 4-Chloro | OH | OH | 6.219 | 3.900 | 0.467 | 2.31 | 2.31 | −0.00 |
| 4-Chloro | OH | H | 5.681 | 3.288 | 0.431 | 2.81 | 2.92 | −0.11 |
| 4-Chloro | H | OH | 5.826 | 3.351 | 0.357 | 3.07 | 2.86 | 0.21 |
| 2,6-Dichloro | OH | OH | 6.647 | 4.470 | 0.840 | 2.37 | 2.49 | −0.12 |
| 2,6-Dichloro | OH | H | 6.109 | 3.858 | 0.804 | 3.04 | 3.09 | −0.05 |
| 2,4-Dichloro | OH | H | 6.092 | 3.724 | 0.726 | 3.35 | 3.28 | 0.07 |
| 2,4-Dichloro | OH | OH | 6.630 | 4.336 | 0.761 | 2.61 | 2.68 | −0.07 |
| 2-Methyl-4-chloro | OH | H | 6.092 | 3.724 | 0.672 | 3.30 | 3.14 | 0.16 |
| 2-Methyl-4-chloro | OH | OH | 6.630 | 4.336 | 0.708 | 2.33 | 2.53 | −0.20 |
| 3-Methyl-4-chloro | OH | OH | 6.630 | 4.471 | 0.932 | 2.90 | 2.69 | 0.21 |
| 3-Methyl-4-chloro | OH | H | 6.092 | 3.859 | 0.896 | 3.30 | 3.30 | 0.00 |
| 2-Methyl-6-chloro | OH | OH | 6.647 | 4.470 | 0.788 | 2.33 | 2.35 | −0.02 |
| 2-Methyl-6-chloro | OH | H | 6.109 | 3.858 | 0.752 | 2.70 | 2.95 | −0.25 |
| 2-Methyl-6-chloro | H | OH | 6.253 | 3.922 | 0.678 | 2.78 | 2.89 | −0.11 |
| 2,6-Dimethyl-4-chloro | OH | OH | 7.041 | 4.700 | 0.913 | 2.76 | 2.89 | −0.13 |
| 2,6-Dimethyl-4-chloro | OH | H | 6.503 | 4.087 | 0.877 | 3.51 | 3.50 | 0.01 |
| 2,6-Dimethyl-4-chloro | H | OH | 6.647 | 4.151 | 0.803 | 3.51 | 3.44 | 0.07 |
| 3,5-Dimethyl-4-chloro | OH | OH | 7.041 | 4.970 | 1.332 | 3.24 | 3.14 | 0.10 |
| 2,4,6-Trimethyl | OH | OH | 7.041 | 4.700 | 0.980 | 3.10 | 3.08 | 0.02 |
| 3,5-Dimethyl-4-chloro | OH | H | 6,503 | 4.358 | 1.296 | 3.68 | 3.74 | −0.06 |
| 2,4,6-Trimethyl | OH | H | 6.503 | 4.087 | 0.944 | 3.47 | 3.68 | −0.21 |
| 3,5-Dimethyl-4-chloro | H | OH | 6.647 | 4.421 | 1.221 | 3.93 | 3.68 | 0.25 |
| 2,6-Dichloro-4-methyl | H | OH | 6.647 | 4.151 | 0.870 | 3.67 | 3.62 | 0.05 |

[a] Based on Eq. 7.

where $r$ is the correlation coefficient, $s$ is the standard error, $F$ is the $F$ statistic, and $n$ is the number of observations.

Both Eqs. 7 and 8 are highly significant in the statistical sense. The $F$ test value exceeds the table value at the 99.5% level: $F\ (3, 25, 0.005) = 5.51$. The Student $t$ values for each coefficient also are very large: 26.9, 37.3, and 80.8 for Eq. 7 and 13.8, 32.3, and 76.7 for Eq. 8. Furthermore, the addition of each variable is highly significant on statistical grounds. The $F$ for the addition of ${}^3\chi_P$ as a second variable in Eq. 10 is significant at the 99.9% level: $F = 29.2 > F\ (1, 25, 0.001) = 13.9$. The $F$ for the addition of ${}^1\chi$ as the third variable in Eq. 7 is equally significant: $F = 67.3 > F\ (1, 24, 0.001) = 14.0$. The $\chi$ indexes for Eq. 7 along with the observed, calculated, and residual values are shown in Table I based on Eq. 7. There are no residuals greater than 1.7 $s$, where $s$ is the standard error of the regression. A plot of observed *versus* calculated values shows no systematic trends in the residuals such as curvature or general increase/decrease as a function of log $1/c$.

To test the integrity of the regression equation, about 15% (four) of the observations were randomly deleted and the three-variable regression was rerun for the remaining 24 observations. The procedure was repeated 10 times, and all of the regression statistics were averaged. None of the regression parameters for the diminished data sets is significantly different from those for the full data set (Table III). This procedure has the effect of testing the predictive power of the equation. When the regression coefficients do not change significantly on deletion of randomly selected sets of observations, one can conclude that the activity of similar compounds will be well predicted also. Furthermore, the average absolute value of the residuals for the deleted observations is 0.15, about the same (0.11) as for the residuals for all 28 observations based on Eq. 7. The range of residuals is also only slightly larger, from −0.34 to +0.30, compared to from −0.25 to +0.25.

Table II—Antiviral Activity for Substituted Benzimidazoles

| | Compound | ${}^6\chi_P$ | Observed | Calculated[a] | Difference |
|---|---|---|---|---|---|
| 1 | Benzimidazole | 0.937 | 2.14 | 2.20 | −0.06 |
| 2 | 2-Methyl | 1.111 | 2.51 | 2.44 | 0.07 |
| 3 | 5-Methyl | 1.236 | 2.72 | 2.62 | 0.10 |
| 4 | 5,6-Dimethyl | 1.452 | 2.72 | 2.92 | −0.20 |
| 5 | 4,6-Dimethyl | 1.468 | 2.82 | 2.94 | −0.12 |
| 6 | 2,5-Dimethyl | 1.438 | 2.89 | 2.90 | −0.01 |
| 7 | 4,5-Dimethyl | 1.394 | 2.96 | 2.84 | 0.12 |
| 8 | 2,5,6-Trimethyl | 1.715 | 3.05 | 3.29 | −0.24 |
| 9 | 2,4,5-Trimethyl | 1.627 | 3.20 | 3.16 | 0.04 |
| 10 | 5,6-Diethyl | 1.912 | 3.39 | 3.56 | −0.17 |
| 11 | 2-Propyl-5-methyl | 1.936 | 3.60 | 3.60 | 0.00 |
| 12 | 2,4,5,6,7-Pentamethyl | 2.051 | 3.66 | 3.76 | −0.10 |
| 13 | 2-Ethyl-5-methyl | 1.757 | 3.74 | 3.35 | 0.39 |
| 14 | 2-Butyl-5-methyl | 2.014 | 3.77 | 3.71 | 0.06 |
| 15 | 2-Isopropyl-5-methyl | 1.979 | 3.77 | 3.66 | 0.11 |

[a] Based on Eq. 9.

**Benzimidazole Antiviral Data Set**—The one-variable correlation with the antiviral data, expressed as the logarithm of concentration to give 75% inhibition of multiplication log $(1/c)$, revealed that the ${}^6\chi_P$ index gives a very good correlation:

$$\log(1/c) = 1.40\ (\pm 0.016)\,{}^6\chi_P + 1.11\ (\pm 0.29) \qquad \text{(Eq. 9)}$$

$$r = 0.950 \quad s = 0.166 \quad F = 120.3 \quad n = 15$$

The statistical significance of the equation is indicated by the fact that the calculated $F$ value exceeds the tabulated value at the 99.5% level: $F = 120.3 > F\ (1, 13, 0.005) = 11.8$.

The addition of ${}^4\chi^v_P$ as a second variable improves the correlation:

$$\log(1/c) = 1.89\ (\pm 0.058)\,{}^6\chi_P - 0.677\ (\pm 0.087)\,{}^4\chi^v_P + 1.04\ (\pm 0.59) \qquad \text{(Eq. 10)}$$

$$r = 0.966 \quad s = 0.144 \quad F = 86.6 \quad n = 15$$

As shown by the appropriate $F$ test, the addition of ${}^4\chi^v_P$ as a second variable is significant above the 95% level: $F = 5.7 > 4.61 = F\ (1, 13, 0.05)$. Observed, calculated, and residual $PC$ values are shown in Table II along with values for ${}^6\chi_P$. Table II is based on Eq. 9.

## DISCUSSION

For both sets of data, the connectivity correlations are of excellent quality with a reasonable number of variables. For the phenyl propyl ethers, the ratio of observations to variables is 28/3 = 9.3; for the benzimidazoles, it is 15/1.

**Phenyl Propyl Ether Antifungal Activity**—For this data set, three variables are required to give a good account of the variation in the biological data. Such a result suggests that more than one structural feature is important in determining the biological activity. If the activity is the result of a generalized dispersion interaction or a simple size effect, the data probably should be described with a single variable, such as ${}^1\chi$ or ${}^1\chi^v$, which has been shown to be important for such effects (4–6).

Equations 1 and 2 account equally well for the experimental data as indicated by their regression statistics. Since the two equations differ only in one variable, ${}^4\chi^v_{PC}$ or ${}^3\chi^v_P$, these two variables probably encode the same information for this data set. Moreover, these two variables should be

Table III—Comparison of Regression Parameters for Diminished and Full Antimicrobial Data Sets for Phenyl Propyl Ethers

| Data Set | Coefficients |  |  | Regression Parameters |  |
|---|---|---|---|---|---|
|  | $^1\chi$ | $^3\chi_P$ | $^4\chi^v_{PC}$ | $r$ | $s$ |
| Diminished[a] | 2.45 (±0.07) | −3.39 (±0.09) | 2.71 (±0.10) | 0.957 (±0.007) | 0.147 (±0.007) |
| Full[b] | 2.44 (±0.09) | −3.29 (±0.09) | 2.71 (±0.03) | 0.957 | 0.149 |

[a] See text for definition. Values are averages of 10 separate runs and in parentheses are given the standard error of the average. [b] From Eq. 7.

highly intercorrelated, an expectation borne out by an examination of the correlation matrix; $^4\chi^v_{PC}$ correlates with $^3\chi^v_P$ with $r = 0.97$. Hence, either equation represents the structure–activity relationship of the data set and may be used for prediction. Subsequent discussion and analysis are limited to Eq. 7 somewhat arbitrarily but also because the $^4\chi^v_{PC}$ index is important for other phenyl ring systems (4, 15).

The two statistically most significant variables in Eq. 7 are $^4\chi^v_{PC}$ and $^3\chi_P$. Whereas $^1\chi$ contains contributions uniformly from the overall molecule, both $^4\chi^v_{PC}$and $^3\chi_P$ depend heavily upon specific branching characteristics and substitution patterns in the molecular skeleton (4). Thus, the spatial arrangement determined by molecular connections is important in determining an optimum interaction; the optimum interaction is reflected by the nature of the $^4\chi^v_{PC}$ and $^3\chi_P$ indexes. Chemical intuition may suggest a possible hydrogen bonding involving the hydroxyl groups at $X$ and/or $Y$. In addition, as subsequent analysis suggests, an interaction involving the substituted phenyl ring, centered at the *para*-position, may be hypothesized.

The two most active compounds (Compounds 25 and 27) are characterized by the largest $^4\chi^v_{PC}$ values, 1.302 and 1.227, respectively. Hence, a search for more active compounds should center on this fact. The negative coefficient on $^3\chi_P$ suggests an unfavorable condition arising from the presence of substituent groupings that lead to large $^3\chi_P$ values. The 1,3,4,5-pattern on the phenyl ring leads to large values for both $^3\chi_P$ and $^4\chi^v_{PC}$. However, the $^4\chi^v_{PC}$ index is increased by the presence of chlorine (and, of course, bromine) whereas the nonvalence (simple connectivity) index $^3\chi_P$ reflects only the atomic arrangement and not the identity of the atoms. The presence of the 1,2-dihydroxy pattern ($X/Y$ occupancy) on the side chain also increases $^3\chi_P$ and decreases the calculated activity.

It would be meaningful if specific structural interpretation could be extracted from the regression equation. This analysis was accomplished by a partitioning on the $^4\chi^v_{PC}$ and $^3\chi_P$ indexes.

As already discussed, each $\chi$ index, $^m\chi_t$, is a summation of subgraph terms, $^mS_j$; the index represents the whole molecule summation of topological features. However, it is possible to factor out of the summation those terms that isolate specific structural features or substructures and to identify those most significantly related to the biological activity. In this manner, the structure–activity relationship study may be focused on selected features of the molecular structure so that significant fragments or substructures may be identified. In this present case, for example, it seems meaningful to explore the substitution pattern in the aromatic ring to determine the arrangement that elicits maximum response.

In this second stage of connectivity analysis, subgraph terms are factored out to explore the specific attributes of structure in the $X/Y$ region of the side chain and the *para*-region of the phenyl ring. The symbol $^m\overline{S}_t$ (the subgraph term symbol with a bar) is used to denote the restricted sum of subgraph terms that encompass only specified portions of the molecular skeleton. For the $X/Y$ region, the symbol $^3\overline{S}_P$ stands for the sum of path-three subgraphs that include the OH at position $X$ or position $Y$ or both. Such subgraphs may be the following molecular fragments: $CH_2CH_2CH_2OH$, $CH_2\underset{|}{C}HCH_2OH$, $CHOHCH_2OH$, and $OCH_2\underset{|}{C}HOH$.

The symbol $^4\overline{S}^v_{PC}$ is defined to contain only path/cluster-four subgraphs containing a *para*-substituent. Some of the possible subgraphs are CHCClCHCH, CHCClC($CH_3$), and > CCClC($CH_3$). To determine the possible significance of these partial $\chi$ indexes, a regression was run in which the $^m\overline{S}_t$ terms were put in place of the complete $^m\chi_t$ indexes. For the equation with $^3\overline{S}_P$, $^4\overline{S}^v_{PC}$, and $^1\chi$, the correlation coefficient is $r = 0.924$. When compared to $r = 0.957$ for Eq. 7, it appears that the partial subgraph terms $^3\overline{S}_P$ and $^4\overline{S}^v_{PC}$ must be capturing the major part of the important structural information in the total indexes $^3\chi_P$ and $^4\chi^v_{PC}$. Furthermore, the correlation strongly suggests that the $X/Y$ region and the ring *para*-region play very important roles in the biological action. That is, the action arises from major interactions centered in these two portions of the molecules rather than from generalized whole molecule effects.

**Benzimidazole Antiviral Activity**—Examination of the regression equations shows that only one variable is required for a very good correlation with the antiviral data. The addition of a second variable does improve the results, $r = 0.962$ compared to $r = 0.950$ (93% variation explained compared to 88%). Either equation can be used for prediction.

In a property–activity relationship study, Hansch found only a modest correlation with the partition coefficient, $r = 0.903$ (16). The alkyl $\pi$ values used in the study are proportional to the number of substituent carbon atoms. Such analysis suggests that only the number of substituent atoms is important and that the substituent size, branching, and pattern of substitution are less important. However, the appearance of high order connectivity path terms, sixth and fourth order, and one valence term in Eqs. 9 and 10 strongly indicates the importance of the alkyl chain length and the pattern of substitution.

To investigate further the question of substitution pattern, the constituent subgraphs were analyzed in a manner somewhat like that described for the phenyl propyl ether data set. In this case, major attention was focused on the efficacy of substitution on the five-membered ring in comparison to the six-membered ring.

All of the subgraphs contributing to the $^6\chi_P$ index are subdivided in various ways. For example, some sixth-order path subgraphs are contained wholly in the benzimidazole nucleus. These subgraphs, given the symbol $R_{56}$, do not include any substituent atoms. Examples are shown in Fig. 1. The symbol $S_6R_6$ stands for the sum of subgraph terms including a substituent on the six-membered ring ($S_6$) and extending only over the six-membered ring ($R_6$). Other subgraphs include a substituent on the six-membered ring but also extend over both the six- and five-membered rings; these substituents contribute to $S_6R_{56}$. In similar fashion, terms may be defined for the five-membered ring as $S_5R_5$ and $S_5R_{56}$. Finally, some sixth-order path subgraphs include substituents on both rings and also extend over parts of both rings; these substituents contribute to $S_6S_5R_{56}$. Thus, for sixth-order paths, there are six subclassifications: $^6S_6R_6$, $^6S_6R_{56}$, $^6S_5R_5$, $^6S_5R_{56}$, $^6S_5S_6R_{56}$, and $^6R_{56}$. The superscript 6 refers to sixth order. An analogous set of subgraph classifications may be obtained for the $^4\chi^v_P$ index as follows: $^4S_6R_6$, $^4S_6R_{56}$, $^4S_5R_5$, $^4S_5R_{56}$, and $^4R_{56}$. Fourth-order paths are not long enough to span the distance involving substituents on both rings. All of these subclassifications may be obtained routinely from the XFUNC computing system that produces the file of $\chi$ indexes.

The first observation to be made is that sixth-order subgraphs contained wholly in the benzimidazole nucleus, $^6R_{56}$, do not contribute much to the correlation: $r^2 = 0.25$. This result is expected since such terms do not vary greatly with substituent changes. Hence, it is expected that subtracting $^6R_{56}$ terms from the total $^6\chi_P$ index would not greatly affect its correlation. The $r$ for $^6\chi_P - {}^6R_{56}$ is 0.92, not a great deal less than the 0.95 found for $^6\chi_P$ alone.

The two best single subgraph terms for correlation with the activity involve five-membered ring substitution: $r = 0.69$ for $^6S_5R_{56}$ and $r = 0.72$

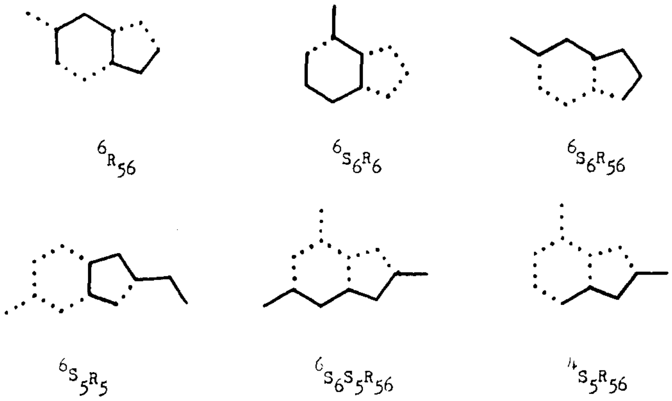


**Figure 1**—*Some constituent subgraphs of $^6\chi_P$ and $^4\chi^v_P$ in alkylbenzimidazoles. Subgraphs are shown in full dark lines; the rest of the graph is in dotted lines.*

for $^4S_5R_5$. These partial $\chi$ indexes, each containing contributions from only a few of the total number of subgraphs, account for about 50% of the variation. This result suggests the importance of substitution on the five-membered ring.

Only six pairs of these subgraph variables yield a correlation equal in quality to Eq. 9, which utilizes all of the subgraphs in $^6\chi_P$; $r \geqslant 0.93$. Three of the six equations contain $^6S_5R_{56}$; the other three contain $^4S_5R_5$. Thus, a partial subgraph term involving five-membered ring substitution is required for quality correlation. No combination of terms that excludes five-membered ring substitution yields a high quality correlation. The best of these two variables is based on $^6S_6R_{56}$ in addition to $^6S_5R_{56}$, for which $r = 0.941$.

These observations indicate the importance of: (*a*) alkyl substitution on the 2-position (five-membered ring) and (*b*) the combination of substitution on both five- and six-membered rings. The compound with the largest value of $^6S_5R_{56}$, 0.6193, is Compound 15, the compound of highest activity.

The subgraph analysis may be continued by examination of three-variable equations involving the $^m\overline{S}_t$ terms. There are 19 sets of three-variable equations with $r \geqslant 0.94$. Sixteen sets include $^6S_5R_{56}$, again underscoring the importance of substitution on the five-membered ring. The other equations contain either $^6S_5R_5$ or $^4S_5R_{56}$. Furthermore, of the seven equations with $r \geqslant 0.95$, all contain $^6S_5R_{56}$. Also prominent in these high quality correlations are $^6S_6R_6$ and $^6S_6R_{56}$. It is significant that the $^6S_6S_5R_{56}$ term does not play an important role in any of the regression equations. Its appearance would suggest that substitution on the 4- and 7-positions is more efficacious than on the 5- and 6-positions. However, there seems to be no discrimination among the six-membered ring positions.

The key term, $^6S_5R_{56}$, is enhanced by large alkyl groups on the 2-position; branched groups produce greater activity. Compare Compound 15 (2-isopropyl) with Compound 11 (2-propyl). Also, Compound 15 (2-isopropyl) is as active as Compound 14 (2-butyl), four carbon atom-substituent unbranched compared to three branched. Apparently, cyclopropyl, *sec*-butyl, and *tert*-butyl groups may enhance activity as substituents on the 2-position.

## CONCLUSIONS

Two general conclusions can be drawn from these studies. First, for both data sets, molecular connectivity provides excellent structure-activity relationship equations; in both cases, these results are superior to previously published Hansch $\pi$, $\sigma$ analyses. Moreover, by systematic elimination of randomly selected phenyl propyl ethers, the significance and reliability of the regression equation were further substantiated. For both data sets, the significant regression variables were partitioned into subgraph terms related to various structural features of the molecules. These partial subgraph terms, $^m\overline{S}_t$, were then entered into regression studies.

For the phenyl propyl ether data set, the results suggest the following conclusions. The principal interactions are related to the *para*-position of the phenyl ring and the *X*/*Y* portion (Table I) of the ether side chain. Substitution leading to increased $^4\chi^v_{PC}$ values augments the activity. Bromine in the *para*-position and methyl in the *meta*-position may increase activity. *vic*-Dihydroxy substitution decreases activity. In addition to these specific structural effects, there is a general whole molecule interaction, perhaps related to dispersion effects.

For the alkylbenzimidazole data set, subgraph analysis shows that substitution of branched (or cyclic) alkyl groups on the 2-position (five-membered ring) is important for high activity. There is no discrimination between the four positions on the six-membered ring.

By use of the partial subgraph terms, $^m\overline{S}_t$, structurally significant portions of the biologically active molecule may be identified in cases where the activity may be highly related to a specific structural feature. The significance of the molecular fragment or substructure may even be put into quantitative terms by using the regression equation for the partial subgraph terms. The method of substructure identification by partial subgraph analysis can now be applied to more complex chemical systems.

## REFERENCES

(1) W. C. Holland, R. L. Klein, and A. H. Briggs, "Introduction to Molecular Pharmacology," Macmillan, New York, N.Y., 1964.

(2) A. Korolkovas, "Essentials of Molecular Pharmacology," Wiley–Interscience, New York, N.Y., 1970.

(3) R. B. Barlow, "Introduction to Chemical Pharmacology," Wiley, New York, N.Y., 1964.

(4) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research," Academic, New York, N.Y., 1976.

(5) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randić, *J. Pharm. Sci.*, **64,** 1971 (1975).

(6) L. B. Kier, W. J. Murray, and L.H. Hall, *J. Med. Chem.*, **18,** 1272 (1975).

(7) L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, **65,** 1806 (1976).

(8) L. H. Hall and L. B. Kier, *ibid.*, **66,** 642 (1977).

(9) T. DiPaolo, L. B. Kier, and L. H. Hall, *Mol. Pharmacol.*, **13,** 31 (1977).

(10) L. B. Kier, T. DiPaolo, and L. H. Hall, *J. Theor. Biol.*, **67,** 585 (1977).

(11) L. B. Kier and L. H. Hall, *Eur. J. Med. Chem.*, **12,** 307 (1977).

(12) F. M. Berger, C. V. Hubbard, and B. J. Ludwig, *Appl. Microbiol.*, **1,** 146 (1953).

(13) C. Hansch and E. J. Lien, *J. Med. Chem.*, **14,** 653 (1971).

(14) I. Tamm, K. Folkers, C. H. Shunk, D. Heyl, and F. L. Horofall, *J. Exp. Med.*, **98,** 245 (1953).

(15) L. B. Kier and L. H. Hall, *J. Med. Chem.*, **12,** 1631 (1977).

(16) "Biological Correlations, The Hansch Approach," W. van Valkenburg, Ed., ACS Advances in Chemistry Series 114, American Chemical Society, Washington, D.C., 1972.

## ACKNOWLEDGMENTS

The authors thank Mr. James D. Fisk and Mr. Philip A. Coy for assistance with the computer calculations. They also appreciate use of the computer centers at Eastern Nazarene College and the Massachusetts College of Pharmacy.