

A Substructural Analysis Method for Structure-Activity Correlation of Heterocyclic Compounds Using Wiswesser Line Notation

GEORGE W. ADAMSON* and DAVID BAWDEN†

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received January 31, 1977

A method of substructural analysis for structure-property correlation of data sets including heterocyclic structures is described. Structural features allowing representation of occurrence and position of heteroatoms, ring fusions, and substituents are derived automatically from Wiswesser Line Notation representations. Structural feature sets of varying degrees of complexity may be derived, suitable either for sets of derivatives of a single ring system or for mixed sets containing several ring systems. The method is evaluated by correlating pK_a values of 169 nitrogen heterocycles, including multisubstituted derivatives of 11 different ring systems. The technique could be carried out automatically with large machine-readable structure-property files and applied to a wide variety of properties.

The investigation of the quantitative relationship between chemical structure and property is currently a field of active investigation, particularly with regard to the development of biologically active compounds.¹⁻³ The methods used have included quantum mechanical calculations⁴ and correlation with physicochemical properties.^{5,6} Empirical methods, using statistical modelling to relate structural features of the compounds under investigation with property, have also been applied,⁷⁻¹⁰ as have pattern recognition techniques,¹¹⁻¹³ and a connectivity index derived from molecular structure has been used to correlate a number of molecular properties.¹⁴ Such empirical methods could be particularly useful if the structural features were derived automatically from the computer-readable structural representations used in chemical information systems,¹⁵⁻¹⁶ since the techniques could then be applied to computer-based files containing both structures and property data.¹⁷⁻¹⁹ An early published example of this methodology, generally known as substructural analysis, used features derived from a fragmentation code,²⁰ and the use of connection tables²¹⁻²³ and Wiswesser Line Notation (WLN)^{24,25} has been demonstrated.

It is of considerable importance, if such methods are to be of wide applicability, especially for the design of biologically active compounds, that they should deal adequately with heterocyclic structures, including multiheteroatomic, multisubstituted, and fused systems, and should allow the investigation of effects due to the relative positions of heteroatoms and substituents. It is also desirable that the methods should be applicable to a variety of ring systems within a single analysis. Similar problems have been encountered in the development of semiempirical Hammett-type linear free energy relationships for heterocyclic systems,²⁶ although these methods have not been applied to large sets of structures containing diverse ring systems.

A recent example of substructural analysis used structural features automatically derived from WLN to represent explicitly positional isomerism and substituent interaction in a correlation of structure with reactivity for a series of polysubstituted benzene derivatives.²⁵ Described below is an extension of the technique to deal with heterocyclic systems and its application to a set of pK_a values for a series of 169 derivatives of 11 nitrogen heterocyclic systems.












DATA

pK_a values measured at 20 °C in aqueous solution for 169 compounds were taken from standard compilations.²⁷ The

* Author to whom correspondence should be addressed.

† Pfizer Central Research, Sandwich, Kent, England.

Table I. Types and Number of Heterocyclic Systems in Analysis

Parent ring system	No. of derivatives	Parent ring system	No. of derivatives
 Pyridine	52	 Cinnoline	5
 Pyridazine	10	 Phthalazine	5
 Pyrimidine	60	 Quinoxaline	5
 Pyrazine	6	 1,2,4-Triazine	2
 Quinoline	13	 1,3,5-Triazine	3
 Isoquinoline	8		

number of derivatives of each of 11 parent ring systems used are shown in Table I. Pyridine derivatives comprise approximately one-third of the set, and pyrimidine derivatives make up another third. The remainder of the set consists of derivatives of seven different ring systems. Just over 20% of the total set comprises derivatives of fused ring systems.

pK_a was regarded as a suitable property value to test this method since it has been measured accurately for derivatives of a variety of heterocyclic ring systems. It has also been recognized as an important factor in determining some kinds of biological activity.^{28,29}

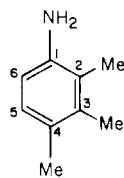
The effects of substituents on pK_a values have been described in detail,²⁷ and comparison of the results of the analyses with these documented effects enabled an assessment of the reliability and usefulness of the technique.

STRUCTURAL FEATURE DERIVATION

An earlier study in the field of substructural analysis derived structural features for benzene derivative from WLN representations, by treating substituent groups as whole units and representing their relative positions in the conventional ortho, meta, and para form.²⁵ The shortest ring path between pairs of substituents is always considered. This simple concept may be extended to heterocyclic structures by treating heteroatoms and ring fusion points as substituents on a parent ring system, in this case a six-membered aromatic system, a method used

in some applications of the Hammett equation.²⁶

An amount of approximation is involved here. Thus in the benzene derivative shown below the *Me-ortho-Me* interactions



2-3 and 3-4 are not identical, because of the different relative position of the NH_2 group. A compromise is necessary between specifying structural features in sufficient detail and keeping the number of variables to an acceptable level. The good correlations achieved in the study of benzene reactivities²⁵ suggests that the level of approximation was not too great for that data set. However the situation is rather different for heterocycles. Thus, for example, it is evident that in the pyridine nucleus (Table II) there are three possible forms of meta interaction: 2-4 (equivalent to 4-6), 3-5, and 2-6. The 2-6 interaction is distinct in that a heteroatom is included between the meta positions. In some of the analyses below, structural features were derived to distinguish between interaction terms involving substituents separated by heteroatoms and those involving only carbon atoms, and their usefulness was assessed.

In general for each set of structures investigated, analyses were carried out on a number of sets of structural features of increasing complexity representing number and type of substituent only or position of substituents relative to heteroatoms, fusion points (if applicable), and/or other substituents. The structures were coded manually in WLN,³⁰ without multipliers or contractions, and structural features were derived by computer program. Details of the structural feature sets used are given in the appropriate sections below.

STATISTICAL ANALYSIS PROCEDURE

These structural features were then correlated with pK_a values by multiple regression analysis,³¹ using a computer manufacturer's statistical analysis package.³² pK_a was assumed to be an additive function of the structural features present, so that its value for the i th structure is given by

$$\text{pK}_{a_i} = \sum_{j=1}^n b_j x_{ij} + \text{constant}$$

where there are a total of n types of structural feature in the set of structures, and x_{ij} is the number of times that the j th feature occurs in the i th structure. The regression coefficient for the j th structural feature, b_j , represents the effect of that substructure in increasing (positive coefficient) or decreasing (negative coefficient) the pK_a values for those compounds in which it occurs. The regression analyses were carried out in a stepwise fashion, with variables being included in order of their pivot elements,³² i.e., approximately in the order of the magnitude of their effect on the variation in the measured pK_a values. The analyses were performed, for the most part, at the 99.99% significance level, i.e., so as to include as many variables as possible. Some, however, were excluded by the program because (i) they were common to all structures; (ii) they had no effect, within the accuracy of the calculation, on pK_a ; or (iii) they were perfectly correlated, i.e., a group of structural features occurred only together in a fixed ratio within the same structures. Only one feature from each perfectly correlated group was included by the program, since only a value for their combined effect can be calculated. In some cases, where a complex structural feature set produced a large number of variables, the analyses were carried out so as to include only those variables significant at 10, 5, or 1%

level. The use of multiple regression analysis enables statistical significance tests to be carried out on overall regression results on the difference between two regression results, on individual regression coefficient values, and on the difference between individual coefficients.^{25,31} The use of the F -test to compare statistically the difference between correlations on the same data using different feature sets is particularly valuable, especially when large numbers of variables are involved. The possibility of good correlations occurring by chance in these circumstances has been noted.³³

Extrapolation of the results of such analyses to predict unknown pK_a values may be carried out by summing the appropriate coefficients.²⁵

The analyses were carried out on the whole set of structures, and on subsets of pyridines, pyrimidines, and the remaining diverse ring systems, to compare the correlations obtained on data sets of varied size and composition.

RESULTS

Pyridine Subset. The subset of 52 pyridine structures was analyzed correlating pK_a with several sets of structural features. These sets included: A, number and type of substituents; B, position of each substituent relative to the heteroatom; C, as B, and additionally including relative positions of each pair of substituents; D, as C, making a distinction between meta substituent pairs in the 2 and 6 positions, separated by heteroatom, and other meta substituent pairs, separated by a carbon atom; E, number and type of substituents, plus relative positions of each substituent pair as in D. Examples of structural feature derivations are given in Table II. The overall results of the regression analyses are shown in Table III. The F values indicate that all the correlations are significant at the 1% level.

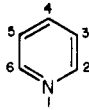
Comparison of the regression results by the F -test indicates that the improved correlations brought about by including firstly positions relative to the heteroatom, and secondly intersubstituent interactions, are both statistically significant at the 1% level. Structural feature set B, including positions relative to the heteroatom but not intersubstituent relative positions, gives a correlation not significantly better than that with set A, including only number and type of substituent. Distinguishing between the two forms of meta interaction did not give a significant improvement in correlation, possibly because of the small number of structures affected. Values for both kinds of meta substituent were available for only two substituent pairs; Me-Me, occurring in nine structures for which both meta interaction terms have negligible coefficient values, and Cl- NH_2 occurring in two structures, for which the 2-6 meta interaction has a negligible value, and other meta interactions show a negative coefficient corresponding to about 0.6 of a pK_a unit.


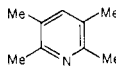
The best correlation for the pyridine subset data is clearly that using structural feature set C, i.e., indicating the position of each substituent relative to the ring nitrogen, and including the relative position of all substituent pairs. The structural features in this set, with their regression coefficients and t statistics are listed in Table IV.

These results are largely interpretable from the electronic properties of the substituent groups, although values were not available in the data set for all the structural features necessary for an exhaustive examination of trends in coefficient values.

Amino, methylamino, and dimethylamino substituents are highly base-strengthening when ortho and para to the heteroatom, while the lesser effect of the methyl group does not appear to be position dependent. Halogen and nitro substituents are base weakening in all positions, while OMe and SMe substituent effects are position dependent, with strongly negative coefficients when ortho to the ring nitrogen, weakly

Table II. Examples of Structural Feature Derivation for Pyridines



Structure (WLN)	Structural feature sets				
	A	B	C	D ^a	E ^a
 (T6NJ DNW)	1 NO ₂	1 NO ₂ -para-RING N	1 NO ₂ -para-RING N	1 NO ₂ -para-RING N	1 NO ₂
 (T6NJ BI CI EI FI)	4 Me	2 Me-ortho-RING N 2 Me-meta-RING N	2 Me-ortho-RING N 2 Me-meta-RING N 2 Me-ortho-Me 2 Me-meta-Me 2 Me-para-Me	2 Me-ortho-RING N 2 Me-meta-RING N 2 Me-ortho-Me 1 Me-meta*-Me ^a 1 Me-meta-Me 2 Me-para-Me	4 Me 2 Me-ortho-Me 1 Me-meta*-Me ^a 1 Me-meta-Me 2 Me-para-Me

^a 2-6 meta interactions denoted as meta*.Table III. Regression Analysis Results for Subset of Pyridine Structures^a

Structural feature set	No. of structural features	No. included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	10	9 + constant	42	0.912	1.15	23.07
B	21	21 + constant	30	0.992	0.41	88.22
C	41	38 + constant	13	0.999	0.15	170.80
D	43	40 + constant	11	0.999	0.12	137.29
E	31	29 + constant	22	0.962	1.06	9.42

^a Number of structures = 52; range of pK_a values = 9.85.

negative when meta, and positive when para. This may be accounted for by the opposed inductive and resonance effects of these groups.²⁷ The most notable coefficient values for the terms representing intersubstituent interactions are those between amino, methylamino, dimethylamino, and methyl groups. All these are base weakening, probably representing inhibition of electron release by similar substituents.³⁴

Pyrimidine Subset. The pyrimidine nucleus (Table V) appears to present more complex problems in deriving structural features useful in structure-property correlation than is the case with pyridine. The terms representing relative position of substituent and heteroatom, in the ortho case, reflect widely different physical situations because of the difference between the 2 and 4 positions, both of which are ortho to a heteroatom. This does not, in fact, affect the overall correlation, since the 4-position ortho interaction is perfectly correlated with the substituent-para-heteroatom term; thus any discrepancy between the two ortho terms will be correlated by the alteration of the "true" value for the para interaction. Substituent position relative to heteroatoms can, therefore, be accounted for equally well by using either explicit position on the ring, or substituent-heteroatom interaction terms. Reliable interpretation of the results is made easier by using explicit position, but at the expense of generality, i.e., of the ability to deal with other classes of compounds simultaneously.

Distinctions between 2-4 and 2-6 meta interactions, including a heteroatom, and 4-6 meta interactions, including a carbon atom, may be made as in the pyridine subset.

The structural feature sets used to analyze the pyrimidine data were very similar to those used for the pyridine structures. They included: F, number and type of substituent; G, explicit substituent positions, i.e., 2, 4 (equivalent to 6), or 5; H, explicit substituent positions plus intersubstituent interaction terms; I, as set H, distinguishing the two forms of meta interaction; J, number and type of substituent, plus relative positions of substituent pairs as in set I. Examples of structural feature derivation for pyrimidines are given in Table V.

The overall results of the regression analyses are shown in Table VI. The *F* values indicate that all the correlations are significant at the 1% level.

Comparison of the correlations by the *F*-test show that inclusion of substituent position, set G, and of relative position of substituents without distinction between the meta interactions, set H, does not significantly improve the correlation beyond that achieved with the structural features representing only number and type of substituent, set F. Only when intersubstituent interaction terms reflecting the two forms of meta interaction are included, i.e., structural feature set I, is the correlation shown to be improved at the 1% significance level. Set J gives a correlation which is not significantly better than that with the simple set F. These results suggest that the positions of substituents relative both to heteroatoms and to other substituents are of importance. Specification of the position of a substituent does not in itself improve the correlation, as it did with the pyridine subset. This may be interpreted as a measure of the nonadditivity of the effects of substituents in the pyrimidine system. The importance of the distinction between the two possible meta relationships between substituents seems to further indicate the importance of intersubstituent effects in this complex system.³⁵

It may be noted that structural feature set I produced 66 variables, i.e., in excess of the number of measured property values. Because of perfect correlation the number of features included was sufficiently small to allow a regression analysis, but with so large a number of variables the predictive usefulness of such a correlation is doubtful, as discussed below. In order to reduce the number of variables, the analysis was repeated in such a way that the regression program omitted variables insignificant at the 10% level. This eliminated approximately half the variables included at the 99.99% level. The structural features included in this 10% level analysis are listed in Table VII, together with their regression coefficients and *t* statistics; the overall result is shown in Table VI. These results again demonstrate the importance of intersubstituent


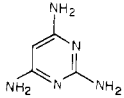
Table IV^a

Structural feature	Regression coefficient	<i>t</i> statistic (13 degrees of freedom)	Perfectly correlated structural features
Me-ortho-RING N	0.71	6.21	
Me-meta-RING N	0.47	4.14	
Me-para-RING N	0.75	5.50	
NH ₂ -ortho-RING N	1.67	9.83	
NH ₂ -meta-RING N	0.69	3.40	
NH ₂ -para-RING N	3.90	24.82	
NHMe-meta-RING N	0.95	2.81	
NHMe-para-RING N	4.37	23.35	
NMe ₂ -para-RING N	4.33	23.18	
OMe-ortho-RING N	-2.01	9.97	
OMe-meta-RING N	-0.41	2.05	
OMe-para-RING N	1.33	6.57	
SMe-ortho-RING N	-1.67	8.29	
SMe-meta-RING N	-0.84	4.18	
SMe-para-RING N	0.68	3.35	
NO ₂ -meta-RING N	-4.52	22.40	
NO ₂ -para-RING N	-3.68	18.26	
Cl-ortho-RING N	-4.13	29.25	Cl-para-NH ₂
Cl-para-RING N	-1.41	7.00	
Br-meta-RING N	-2.38	11.80	
Br-para-RING N	-1.47	7.30	
Me-ortho-Me	0.08	0.65	
Me-meta-Me	0.01	0.11	
Me-para-Me	0.02	0.18	
Me-ortho-NH ₂	-0.30	2.24	
Me-meta-NH ₂	-0.31	2.09	
Me-ortho-NHMe	-0.32	2.21	
Me-meta-NHMe	-0.78	5.22	
Me-ortho-NMe ₂	-1.27	8.68	
NH ₂ -ortho-NH ₂	-0.73	3.10	
NH ₂ -para-NH ₂	-1.10	3.98	
NH ₂ -ortho-NHMe	-0.78	2.71	
NO ₂ -ortho-NH ₂	0.16	0.64	
NO ₂ -ortho-NHMe	0.05	0.19	
Cl-ortho-NH ₂	0.29	0.89	Cl-meta-Cl
Cl-meta-NH ₂	Excluded by regression program		
Br-ortho-NH ₂	0.23	0.85	
Br-ortho-NHMe	0.19	0.67	
Br-ortho-NMe ₂	-0.72	2.53	
Regression constant	5.29	40.38	

^a Pyridine subset, structural feature set C.

interaction in this system; of the 30 structural features significant at the 10% level, half are substituent pair relative positions, and all except one of these represent meta interaction. The major trends in these results appear reasonable

Table V. Examples of Structural Feature Derivation for Pyrimidines

Structure (WLN)	Structural feature sets				
	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i> ^a	<i>J</i> ^a
 (T6N CNJ DZ)	1 NH ₂	1 4-NH ₂	1 4-NH ₂	1 4-NH ₂	1 NH ₂
 (T6N CNJ BZ DZ FZ)	3 NH ₂	1 2-NH ₂ 2 4-NH ₂	2 2-NH ₂ 2 4-NH ₂ 3 NH ₂ -meta-NH ₂	1 2-NH ₂ 2 4-NH ₂ 1 NH ₂ -meta-NH ₂ 2 NH ₂ -meta*-NH ₂	3 NH ₂ 1 NH ₂ -meta-NH ₂ 2 NH ₂ -meta*-NH ₂

^a 4 and 6 positions are equivalent and denoted as 4; meta* denotes 2-4 and 2-6 meta interactions.

on the basis of known electronic factors.^{27,35} Amino derivative substituents in the 2 and 4 positions increase pK_a markedly, while halogen and nitro groups reduce pK_a , particularly in the 5 position. OMe and SMe substituents exert a positive effect on pK_a in the 4 position and a negative effect in the 2 position. Of the substituent interaction terms amino substituents meta to one another show a negative effect on pK_a , reduced when the substituents are separated by a heteroatom, and presumably indicating a mutual inhibition factor. Amino substituents meta to halogens, OMe, and SMe show a strong interaction reducing pK_a , except when the amino substituent is separated by a heteroatom from OMe, when a positive effect on pK_a is observed.

When these results are compared with the coefficients for the same structural feature set, analyzed at the 99.99% level, it appears that the same general conclusions may be drawn from each. The 10% level analysis, including only half the number of structural features, is less complex and hence easier to interpret; the 99.99% level analysis, on the other hand, allows for the observation of more trends in individually insignificant coefficients. The results of the two analyses are not in total accord; thus the Br-ortho-OMe structural feature which has a significant positive coefficient in the 10% level analysis appears to have a negligible effect in the analysis involving more variables. Some variation of this kind in particular coefficient values in different analyses is to be expected, and acts as a caution against placing undue emphasis on single coefficients in interpreting the results of such correlations.

Diverse Subset. In order to deal with this subset of derivatives of nine ring systems the approximations introduced by the use of structural features representing the simplest form of relative position were accepted. To use, for example, explicit substituent positions or more accurately specified interaction terms as applied to particular ring systems would have introduced so many variables as to defeat the purpose of a generalized analysis.

Four sets of structural features were used including K, number and type of substituents only (including heteroatoms and fused rings); L, positions of substituents (including ring fusion points) relative to heteroatoms; M, as L, plus positions of substituents relative to ring fusion points; N, as M, plus positions of substituents in relation to one another. Examples of structural feature derivation are given in Table VIII.

The overall results of the regression analyses are given in Table IX. All the correlations are significant at the 1% level.

Comparison of the regression by the *F*-test shows that, compared with the analysis using set K structural features, no significant improvement is brought about by the use of structural feature set L, and an improvement only at the 10%

Table VI. Regression Analysis Results for Subset of Pyrimidine Structures^a

Structural feature set	No. of structural features	No. included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
F	10	10	50	0.983 ^b	0.98	143.32 ^b
G	21	21 + constant	38	0.934	0.96	12.37
H	56	47 + constant	12	0.974	1.08	4.72
I (99.99% level)	66	53 + constant	6	0.999	0.08	113.12
I (10% level)	66	30 + constant	29	0.993	0.36	68.32
J	53	47 + constant	12	0.984	0.84	7.79

^a Number of structures = 60; range of pK_a values = 8.73. ^b Correlation coefficient and hence F value are relatively high owing to the lack of a regression constant.

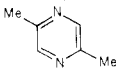
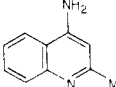
Table VII^a

Structural feature ^b	Regression coefficient	t statistic (29 degrees of freedom)	Perfectly correlated features
2 Me	0.76	1.79	Me-meta*-Me
2 NH ₂	2.23	9.98	
4 NH ₂	4.09	21.90	
2 NHMe	2.53	6.72	
4 NHMe	2.14	11.66	
2 NMe ₂	2.53	8.84	
4 NMe ₂	4.56	17.14	
2 OMe	-0.91	3.22	
2 SMe	-0.90	3.65	
4 SMe	0.73	2.95	
2 Cl	-1.08	2.69	
4 Cl	-1.42	4.45	
5 Cl	-1.76	5.65	
5 Br	-1.83	8.38	
5 NO ₂	-3.95	18.73	
NH ₂ -meta*-NH ₂	-0.81	3.66	
NH ₂ -meta*-NMe ₂	-0.67	1.83	
NH ₂ -meta-NH ₂	-3.96	12.62	
NH ₂ -meta-NHMe	-1.68	5.55	
NHMe-meta-NMe ₂	-2.08	4.45	
NH ₂ -meta*-OMe	1.03	3.15	
NMe ₂ -meta*-OMe	1.17	3.33	
NH ₂ -meta-OMe	-2.33	7.23	
NMe ₂ -meta-OMe	-2.08	4.61	
NH ₂ -meta-SMe	-2.60	7.07	
NMe ₂ -meta-SMe	-2.49	5.10	
NH ₂ -meta-Cl	-2.42	6.50	
NMe ₂ -meta-Cl	-2.48	4.62	
Br-ortho-OMe	-1.41	3.18	
Me-meta*-NH ₂	0.37	2.05	
Regression constant	1.77	10.96	

^a Pyrimidine subset; structural feature set I; analysis at 10% level. ^b meta* denotes 4-6 meta interaction; 4 denotes either 4 or 6 position.

level by the use of set M. The analysis with set N structural features brings about no further significant improvement.

Table VIII. Examples of Structural Feature Derivation for Diverse Structures

Structure (WLN)	Structural feature sets			
	K	L	M	N
 (T6N DNJ B1 E1)	2 RING N 2 Me	RING N-para-N 2 Me-ortho-RING N 2 Me-meta-RING N	RING N-para-N 2 Me-ortho-RING N 2 Me-meta-RING N	RING N-para-N 2 Me-ortho-RING N 2 Me-meta-RING N
 (T66 BNJ C1 EZ)	1 RING N 1 FUSED RING 1 NH ₂ 1 Me	1 Me-ortho RING N 1 NH ₂ -para-RING N 1 RINGFUSION-ortho-RING N 1 RINGFUSION-meta-RING N	1 Me-ortho-RING N 1 NH ₂ -para-RING N 1 RINGFUSION-ortho-RING N 1 RINGFUSION-meta-RING N 1 Me-meta-RINGFUSION 1 Me-para-RINGFUSION 1 NH ₂ -ortho-RINGFUSION 1 NH ₂ -meta-RINGFUSION	1 Me-para-Me 1 Me-ortho-RING N 1 NH ₂ -para-RING N 1 RINGFUSION - ortho-RING N 1 RINGFUSION-meta-RING N 1 Me-meta-RINGFUSION 1 Me-para-RINGFUSION 1 NH ₂ -ortho-RINGFUSION 1 NH ₂ -meta-RINGFUSION 1 Me-meta-NH ₂

In view of the known general similarity of substituent effects in a variety of nitrogen heterocyclic systems,²⁷ this suggests that the approximations involved in the use of simple relative position terms for groups of diverse structures are too great to allow highly significantly improved calculations.

Total Set. The subsets of pyridines, pyrimidines, and diverse structures were combined, giving a total of 169 structures. This set was analyzed using the same types of structural features as for the subset of diverse structures, so as to allow for the different ring systems in this combined group of structures.

The sets of structural features used were as follows: O, number and type of substituent only (including heteroatoms and fused rings); P, positions of substituents (including ring fusion points) relative to heteroatoms; Q, as P, plus positions of substituents relative to ring fusion points; R, as Q, plus positions of substituents in relation to one another.

The results of the regression analyses are summarized in Table X. The correlations with structural feature sets P and Q are not significantly better than that with set O, and the analysis with set R is superior only at the 10% level (all the analyses being carried out so as to include as many features as possible). An analysis with set R, excluding features insignificant at the 10% level, was significantly better at 5% than the analysis using structural feature set O.

Structural features, with regression coefficients, etc., for the analyses with set O structural features and with set R structural features at the 10% level are listed in Tables XI and XII, respectively.

The results from the set O analysis reflect accurately at a simple level the effects of substituents upon pK_a , which are largely as might be expected from the known electronic properties of these substituents.²⁷ Thus ring nitrogen atoms have a negative coefficient, reflecting the base-weakening effect of multiheteroatomic substitution. Amino, methylamino, and dimethylamino substituents are base strengthening, as is the methyl group to a lesser extent, while halogens and nitro substituents are strongly base weakening. OMe and SMe substituents show a weak effect, due to an averaging out of

Table IX. Regression Analysis Results for Subset of Diverse Structures^a

Structural feature set	No. of structural features	No. included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
K	7	7 + constant	49	0.817	1.35	14.05
L	19	18 + constant	38	0.878	1.28	7.10
M	33	26 + constant	30	0.934	1.08	7.89
N	41	34 + constant	22	0.955	1.04	6.71

^a Number of structures = 57; range of pK_a values = 9.16.**Table X.** Regression Analysis Results for Total Set of Structures^a

Structural feature set	No. of structural feature	No. included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
O	11	11 + constant	157	0.884	1.22	51.04
P	30	29 + constant	139	0.891	1.26	18.46
Q	44	38 + constant	130	0.905	1.22	15.48
R (99.99% level)	87	78 + constant	90	0.949	1.09	10.45
R (10% level)	87	26 + constant	142	0.930	1.01	34.96

^a Number of structures = 169; range of pK_a values = 11.35.**Table XI^a**

Structural feature	Regression coefficient	t statistic ^a
RING N	-2.58	13.94
Me	0.72	5.09
NH ₂	1.75	11.89
NHMe	1.98	7.38
NMe ₂	2.85	8.14
OMe	-0.42	1.47
SMe	-0.58	1.81
Cl	-2.75	8.03
Br	-1.4	3.32
NO ₂	-3.44	8.07
Fused ring	-0.16	0.60
Regression Constant	7.85	20.40

^a 157 degrees of freedom. ^a Total set; structural feature set O.

their contributions at different positions observed in the more detailed analyses described above. The presence of a fused ring has a small overall effect upon pK_a , which may again be partly due to an averaging effect.

The results listed in Table XII show some aspects of the effects of substituents upon pK_a in finer detail. Thus the effect of relative position of heteroatoms upon the base-weakening effect of the introduction of more than one ring nitrogen is shown, as is the base strengthening due to amino-type substituents ortho and para to the heteroatom. For the other substituent groups the position relative to ring nitrogen for their more reliably assessed effects are shown, for example, the negative coefficients of the OMe and SMe groups in the ortho position, probably due to a large inductive effect. The intersubstituent interaction terms included, and therefore statistically significant for the whole set, are almost entirely those noted in the pyrimidine subset, emphasizing the importance of substituent interaction in the pyrimidine system. The term representing ring fusion has only a relatively small negative coefficient, again showing the small effect of ring fusion on pK_a , while the only structural feature included representing interaction between substituent and fused ring demonstrates the base-strengthening effect of the amino group ortho to a ring fusion.

DISCUSSION

The work described above shows that substructural analysis techniques can be used to produce structural features which allow detailed analysis of data sets containing heterocyclic compounds. It further indicates that such analyses can produce

Table XII^a

Structural feature	Regression coefficient	t statistic (142 degrees of freedom)	Perfectly correlated structural features
RING N-ortho-RING N	-1.41	5.44	
RING N-meta-RING N	-3.24	12.35	
RING N-para-RING N	-3.43	10.70	
NH ₂ -ortho-RING N	1.49	8.63	
NH ₂ -para-RING N	2.69	11.48	
NHMe-para-RING N	3.21	9.12	
NMe ₂ -ortho-RING N	1.66	5.63	
NMe ₂ -para-RING N	2.84	7.76	
Me-ortho-RING N	0.54	3.57	
Me-meta-RING N	0.59	3.57	
Cl-ortho-RING N	-2.95	6.25	
Cl-para-RING N	-1.01	1.99	
Br-meta-RING N	-1.28	5.29	
OMe-ortho-RING N	-0.60	2.76	
SMe-ortho-RING N	-0.74	3.18	
NO ₂ -meta-RING N	-2.46	10.22	
NO ₂ -para-RING N	-3.52	3.44	
NH ₂ -meta-NH ₂	-2.41	6.51	
NH ₂ -meta-NHMe	-2.41	3.86	
NH ₂ -meta-NMe ₂	-1.59	1.87	
NHMe-meta-NHMe	-1.16	1.90	NHMe-ortho-NHMe
NHMe-meta-NMe ₂	-3.22	2.85	
Cl-meta-NHMe	3.39	4.45	
Br-ortho-OMe	2.61	2.30	
RING N-meta-RING FUSION	-0.31	1.94	
NH ₂ -ortho-RING FUSION	1.67	3.87	
Regression constant	5.13	31.00	

^a Total set structural feature set R; analysis at 10% level.

reliable and potentially useful results.

It has shown that a more detailed analysis is possible for a set of derivatives of a common parent ring system than for a set of diverse ring systems. An inverse relationship exists between the generality of application of a substructural analysis technique of this kind and the specificity of the structural features which may be derived. Although it is advantageous from this point of view to deal only with closely related structures at one time, it has been demonstrated that an analysis in considerable depth for derivatives of a number of ring systems is possible, using simple procedures, although it may often be the case, as in the examples here, that highly significant improvements in correlation are not brought about. A more complex procedure for structural feature derivation,

perhaps based on a minimum spanning tree algorithm, could be useful in making possible further generalization of such analyses. Alternatively ring systems could be fragmented into individual rings, or still smaller units, by relatively simple procedures.³⁶

From the analyses on this data set, and related work,³⁷ it appears that in many cases significantly better correlations are to be obtained by using the more detailed structural feature sets, in this case those including interaction terms. This raises the problem of the very large number of variables which may be included in such analyses. One potentially useful method of reducing the number of variables is to carry out the regression analysis in such a way that variables insignificant at a particular confidence level are omitted from the calculation. This has been demonstrated in the work above, using the 10% significance level. If the aim of the analysis is the investigation, perhaps in a qualitative sense, of the factors involved in a structure-property relationship, it may be advantageous to use the analysis of highest statistical significance, even if this contains a very large number of variables. Useful information may be gained in this way, provided that the results are interpreted in terms of trends among the coefficients, which may individually be statistically insignificant. Analyses carried out with omission of less significant variables may be useful in that they concentrate on the more important factors.

If the aim of the analysis is quantitative prediction of unknown property values, the complex structural feature sets, even if giving significantly better correlations, may be of limited use. Because of the greater specificity of structural feature description in complex sets, it is likely that in many cases coefficient values for the features in a structure not included in the analysis will not be available, either because these structural features do not occur in any structure in the analyzed set, or because they are perfectly correlated with other structural features. The extent of perfect correlation almost always increases with increasing complexity of structural feature sets. The simpler feature sets, with which this problem is less likely to arise, may be more useful for predictive purposes.

It should be noted that, although only multiple regression analysis has been used in this work, the type of structural features derived here could be used to enable heterocyclic structures to be dealt with by other statistical methods, for example, cluster analysis²² or pattern recognition techniques.¹¹⁻¹³

Substructural analysis techniques of this kind, in common with other additive modelling procedures, do not partition the effect of substructures to electronic, steric, lipophilic, and similar factors, as is done by the semiempirical techniques.⁶ From the results above, and those in other cited work, it is evident that this does not detract from the usefulness of these methods for achieving potentially useful empirical correlations. The coefficient values from such analyses may then be interpreted in physico-chemical terms. It is possible that a major use of substructural analysis methods could be as a first-stage approach, to identify major factors in the structure-activity relationship and to aid subsequent more detailed analysis by other methods.

A technique such as the one described above would be particularly suitable for this purpose, since it involves the widely used WLN structure representation, is simple and computationally economical, and is applicable to a wide variety of properties and structural type. It should therefore be capable of providing useful results to workers in a number of different fields.

EXPERIMENTAL

The programs were run on the University of Sheffield 1907

computer. The WLN fragmentation program was written in ICL COBOL and required 12K of core storage with CPU times ≤ 90 s for the set of 169 structures.

The WLN strings representing substituents which were not fragmented into smaller units and terms representing heteroatoms and ring fusion points were stored with their locants, and relative positions were derived by examination of each locant pair. The procedure is applicable to any structure which may be regarded as a substituted six-membered aromatic ring. Other ring systems could be included by utilizing an appropriate dictionary of locant pairs or, more generally, by a spanning-tree algorithm.

The multiple regression analyses were performed using the ICL statistical analysis package. Core storage required was ≤ 32 K and CPU times ≤ 135 s.

ACKNOWLEDGMENT

We thank Professor M. F. Lynch, Dr. G. E. Vleduts, Miss J. A. Bush, and Mr. P. Willett for valuable discussions, and the Department of Education and Science (London) for the award of a Postgraduate Research Studentship to D. Bawden.

LITERATURE CITED

- (1) G. Redl, R. D. Cramer, and C. E. Berkhoff, "Quantitative Drug Design", *Chem. Soc. Rev.*, **3**, 273-292 (1974).
- (2) P. N. Craig, "Structure/Property Correlations", in "Chemical Information Systems", J. E. Ash and E. Hyde, Ed., Ellis Horwood, Chichester, 1975, p 259.
- (3) F. D. Kover, "Structure-Activity Correlation Bibliography", NTIS Report PB-240 658, 1975.
- (4) L. B. Kier, "Molecular Orbital Theory in Drug Research", Academic Press, New York, N.Y., 1971.
- (5) M. S. Tute, "Principles and Practice of Hansch Analysis", *Adv. Drug Res.*, **6**, 1-77 (1971).
- (6) A. Verloop, "The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies", in "Drug Design", Vol. 3, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1972, p 133.
- (7) T. L. Bruice, N. Kharasch, and R. J. Winzler, "A Correlation of Thyroxine-Like Activity and Chemical Structure", *Arch. Biochem. Biophys.*, **62**, 305-317 (1956).
- (8) S. M. Free and J. W. Wilson, "A Mathematical Contribution to Structure-Activity Studies", *J. Med. Chem.*, **7**, 395-399 (1964).
- (9) K. Bocek, J. Kopecky, M. Krivucova, and D. Vlachova, "Chemical Structure and Biological Activity of p-Disubstituted Derivatives of Benzene", *Experientia*, **20**, 667-668 (1964).
- (10) J. Kopecky, K. Bocek, and D. Vlachova, "Chemical Structure and Biological Activity of m- and p-Disubstituted Derivatives of Benzene", *Nature (London)*, **207**, 981 (1965).
- (11) B. R. Kowalski and C. F. Bender, "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 775 Biological Activity Test", *J. Am. Chem. Soc.*, **96**, 916-918 (1974).
- (12) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, and R. I. Geran, "Pattern Recognition and Structure-Activity Relationship Studies", *J. Med. Chem.*, **18**, 539-545 (1975).
- (13) A. J. Stuper and P. C. Jurs, "Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques", *J. Am. Chem. Soc.*, **97**, 182-187 (1975).
- (14) L. B. Kier, W. J. Murray, and L. H. Hall, "Molecular Connectivity. 4. Relationships to Biological Activities", *J. Med. Chem.*, **18**, 1272-1274 (1975).
- (15) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", Macdonald-Elsevier, London-New York, 1971.
- (16) J. E. Ash and E. Hyde, Ed., "Chemical Information Systems", Ellis Horwood, Chichester, 1975.
- (17) V. B. Bond, C. M. Bowman, N. L. Lee, D. R. Peterson, and M. H. Reslock, "Interactive Searching of a Structure and Biological Activity File", *J. Chem. Doc.*, **11**, 168-170 (1971).
- (18) C. Hansch, A. Leo, and D. Elkins, "Computerized Management of Structure-Activity Data. 1. Multivariate Analysis of Biological Data", *J. Chem. Doc.*, **14**, 57-61 (1974).
- (19) M. A. Oxman, H. M. Kissman, J. M. Burnside, J. R. Edge, C. B. Haberman, and A. A. Wykes, "The Toxicology Data Bank", *J. Chem. Inf. Comput. Sci.*, **16**, 19-21 (1976).
- (20) R. D. Cramer, G. Redl, and C. E. Berkhoff, "Substructural Analysis. A Novel Approach to the Problem of Drug Design", *J. Med. Chem.*, **17**, 533-535 (1974).
- (21) G. W. Adamson and J. A. Bush, "Method for Relating the Structure and Properties of Chemical Compounds", *Nature (London)*, **248**, 406-408 (1974).
- (22) G. W. Adamson and J. A. Bush, "A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic

- Classification of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **15**, 55-58 (1975).
- (23) G. W. Adamson and J. A. Bush, "Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc., Perkin Trans. 1*, 168-172 (1976).
- (24) G. W. Adamson and D. Bawden, "A Method of Structure-Activity Correlation Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **15**, 215-220 (1975).
- (25) G. W. Adamson and D. Bawden, "An Empirical Method of Structure-Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **16**, 161 (1976).
- (26) O. Exner in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, London, 1972, pp 44-46.
- (27) A. Albert in "Physical Methods in Heterocyclic Chemistry", A. R. Katritzky, Ed., Academic Press, London: Vol. 1, 1963, p 2; Vol. 3, 1971, p 1.
- (28) T. Fujita, "The Analysis of Physiological Activity of Substituted Phenols with Substituent Constants", *J. Med. Chem.*, **9**, 797-803 (1966).
- (29) J. P. Tollenaere, "Structure-Activity Relationships of Three Groups of Uncouplers of Oxidative Phosphorylation: Salicylanilides, 2-Tri-fluoromethylbenzimidazoles, and Phenols", *J. Med. Chem.*, **16**, 791-796 (1973).
- (30) E. G. Smith and P. A. Baker, "The Wiswesser Line-Formula Chemical Notation", 3rd ed, Chemical Information Management Inc., Cherry Hill, N.J., 1976.
- (31) G. W. Snedecor and W. G. Cochran, "Statistical Methods", 6th ed, Iowa State University Press, Ames, Iowa, 1967.
- (32) "Statistical Analysis Mark 2 Applications Package", ICL Technical Publication 4301, International Computers Ltd., London, 1971.
- (33) J. G. Topliss and R. J. Costello, "Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis", *J. Med. Chem.*, **15**, 1066-1068 (1972).
- (34) J. E. Dubois, J. J. Aaron, O. Alcais, J. P. Doucet, F. Rothenberg, and R. Ucan, "A Quantitative Study of Substituent Interactions in Aromatic Electrophilic Substitution. 1. Bromination of Polysubstituted Benzenes", *J. Am. Chem. Soc.*, **94**, 6823-6828 (1972).
- (35) B. Roth and J. Z. Strelitz, "The Protonation of 2,4 Diaminopyrimidines. 1. Dissociation Constants and Substituent Effects", *J. Org. Chem.*, **34**, 821-836 (1969).
- (36) P. Willett, M.Sc dissertation, PGSLS, Sheffield University, 1976.
- (37) D. Bawden, Ph.D Thesis (in preparation), Sheffield University, 1976.

On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism†

MILAN RANDIĆ

Energy & Mineral Resources Research Institute, Iowa State University, Ames, Iowa 50011

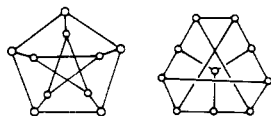
Received November 8, 1976

Use of a canonical numbering based on a particular interpretation of the adjacency matrix of a graph is advocated. The proposed numbering may be used in tests of isomorphism of graphs (molecular skeletons). It also makes possible study of the symmetry properties of graphs and recognition of equivalent vertices. The adjacency matrices based on the proposed canonical numbering may serve as a basis for ordering structures in a sequence. The relative position in the sequence is determined by the magnitude of the associated binary code derived by reading the entries of the adjacency matrix row by row from left and right and from top to bottom. For selected structures, such as isomers of paraffins, the derived ordering parallels certain molecular properties and thereby points to a topological origin of some correlations. This also suggests that the particular numbering involves some inherent features of the connectivity in molecular skeletons and graphs. The particular numbering hence may provide a basis for a systematic nomenclature which does not require supplementary rules. Although ultimately any labeling scheme is arbitrary, it is argued that the scheme proposed has additional properties which will facilitate solving some problems associated with graphs, and hence deserves some attention.

INTRODUCTION

Numbering of the atoms in a molecule and the related problem of numbering of vertices in a graph, apart from their apparent use in chemical documentation and nomenclature, have importance in other areas of science. In chemistry, chemical physics, statistical mechanics, and the theory of disordered structures, graphs provide a convenient representation of the combinatorial possibilities, facilitate visualizing contributing terms in expansions, and may also lead to a structural interpretation of correlation parameters.

One of the most important problems in the study of graphs is that of recognizing identical graphs.¹ It is generally quite difficult and tedious to assert that two graphs have an identical connectivity, as is well illustrated with well-known distinctive representations of the following Petersen graph:



A trial-and-error matching is time consuming and may require checking all $n!$ possibilities. This is essentially implemented

in the node-to-node search,² an early systematic attempt to resolve the problem. The scheme requires extensive book-keeping of the examined possibilities, involves considerable backtracking, and becomes impractical for applications to graphs of medium size and complexity. Alternative schemes considered in the literature attempt to recognize a singular property or a selection of properties that could differentiate between nonisomorphic graphs. Such properties include the graph spectrum,³ distribution of valences of vertices and discrimination of edge types, and examination of subgraphs and their characteristic polynomials.⁴ Sussenguth listed a set of crucial qualities;⁵ however, if one uses any combination of criteria such as mentioned above, one cannot be sure that all pairs of graphs would be differentiated by the criteria. This ultimate uncertainty undermines such efforts, which at best provide a list of necessary conditions for isomorphism.

The problem of graph isomorphism is related to the question of ordering graphs in a sequence. A complete order relation on graphs establishes which of two selected graphs precedes the other. So a search for criteria for ordering of graphs is fundamentally the same problem as the search for isomorphism. It is generally recognized that the use of a standard numbering procedure for vertices makes the problem of establishing isomorphism in graphs trivial and, one could add, makes ordering of graphs straightforward. The problem is in designing such a scheme which will apply to any graph so that the rules give a unique numbering; for a scheme to be practical,

† Portions of this material were presented to the Division of Chemical Information, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 30, 1976, under the title "Symmetry Properties of Graphs" within the Symposium on Application of Nonnumerical Mathematics in Chemistry.