STEREOCHEMISTRY OF HETEROCYCLES

XXXIII.* ANALYSIS OF THE IR SPECTRA OF STEREOISOMERIC SUBSTITUTED 1,3-DIOXANES BY MEANS OF A COMPUTER

G. N. Vostrov, A. I. Gren',
A. V. Bogat-skii, I. B. Gernega,
V. P. Teptya, and Yu. P. Adler

UDC 547.841:543.422.4:541.634

The possibility of the use of the method of correlation pleiads for the isolation of the absorption regions in the IR spectra of substituted 1,3-dioxanes in which the characteristic absorption bands for the individual stereoisomers of the substances appear was studied. It is shown that the configurations of the investigated compounds can be determined on the basis of the characteristic absorption bands by instructing a computer to distinguish forms.

The characteristic absorption bands of the cis or trans isomers in the IR spectra of stereoisomeric substituted 1,3-dioxanes [2] are found at 600-700 cm^{-1}. It has been found to be impossible to find regular changes in the IR spectra of geometrical isomers in the shortwave region when the usual methods of spectral analysis are used.

The present communication is devoted to a study of the possibility of the use of the theory of recognition of forms by means of a computer for the analysis of the IR spectra of cis and trans isomers of substituted 1,3-dioxanes and for finding the most informative absorption bands for the individual stereoisomers. We think that the mathematical apparatus of this method may prove to be useful in investigations of this sort.

The set of absorption bands in the IR spectra of these compounds was used as the starting information regarding the cis and trans isomers of 2,5-dialkyl-, 2,5,5-trialkyl-, 4,5-dialkyl-, and 2,5-dialkyl-5-alkoxyalkyl-1,3-dioxanes, the configurations of which were proved previously by PMR spectroscopy (for example, see [3]).

In the first step we evaluated the informative character of the elements of this description in order to select the wave numbers containing the greatest amount of information regarding the molecular configuration. We simultaneously investigated the correlation relationships between the most informative absorption bands. We used the method of correlation pleiads [4] to evaluate the correlation dependences between the wave numbers and the groupings of the latter. This made it possible to isolate groups of wave numbers that are intimately correlated with one another or are not correlated by are close in value. The existence of groups of the first type can be explained by the fact that in the series of substituted 1,3-dioxanes under consideration these bands have an identical tendency to undergo change as a function of the type of isomer. The existence of the second group follows from the assumption that the absorption bands of the individual elements of the structure are shifted on passing from one compound to another. In both cases, each such group of wave numbers can be represented by one wave number.

The primary information on the informative character of the individual absorption bands was represented as a matrix with absorption bands with spacings of 10-12 cm^{-1} corresponding to its columns and in-

* See [1] for communication XXXII.

I. I. Mechnikov Odessa State University. Odessa Polytechnic Institute. Translated from Khimiya Geterotsiklicheskikh Soedinenii, No. 9, pp. 1189-1195, September, 1974. Original article submitted September 24, 1973.

dividual compounds corresponding to its rows. The entire absorption region was thus broken down into 50 intervals. In this case, all of the absorption bands present in the spectrum for concrete compounds were recorded in the corresponding intervals, while the remaining intervals were filled with zeros.

Let us now imagine that a certain set of random values x is fixed. We will isolate the maximum and minimum values, and we will divide the interval between them into k parts. We next construct k vectors $\bar{y}_1 = (y_{11}, \ldots, y_{1N})$ and $\bar{y}_k = (y_{k1}, \ldots, y_{kN})$, which correspond to these intervals. For this, we will take x_1 as the first component of that vector in the interval of which it lies, and we will set the first components of the remaining vectors equal to zero. We then deal with the remaining random x values in the same way.

We note that the correlative character of random values $y_1, \ldots, y_k$ will be close to zero in the general case.

An analysis of the matrix of the correlative character of the absorption bands showed that precisely the adjacent absorption bands practically do not correlate. This is not difficult to understand if one takes into account the fact that absorption bands of this sort are most frequently a shifting of certain vibrations associated with the appearance in the molecule of some new structural element. Hence, artificial breakdown of the entire absorption region into narrow intervals may distort the picture. To eliminate this, we used the method indicated above to isolate the absorption regions corresponding to the correlated intervals, which are also correlated wave numbers.

The set of correlated signs of the groups will give a simplified description of the objects under investigation, in this case the cis and trans isomers of 2,5-substituted 1,3-dioxanes, with minimal losses of information.

To evaluate the informative character of the indexes (the wave numbers) of the isomers we use the following expressions:

$$S(i) = \sum_{j=1}^{n} |\rho_{ij}|; \quad R(i) = \sum_{j=1}^{n} \rho_{ij}, \quad i = (1, \ldots, n) \ , \tag{1}$$

where ρ_{ij} is the coefficient of correlation between the i-th and j-th wave numbers, and n is the number of indices [4, 5]. Those wave numbers that are the form of the group with a high level of correlation in absolute magnitude are isolated in the first expression as the most informative wave numbers. Indices, the corresponding groups of which contain frequencies that are correlated in pairs primarily with positive correlation coefficients, are preferred in the second expression. These groups are more uniform from the point of view of the character of the linear dependence between them. The S(i) and R(i) functions can be considered to be a measure of the informational value of the i-th wave numbers. The results of treatment of the matrix of the starting data and the S(i) and R(i) values are presented in Table 1.

In conformity with the algorithm for arrangement of the indices in the order of decreasing informative character, the index for which S(i) takes on the greatest value is isolated first. The R matrix was then converted by using the expression of partial correlation [6] in order to eliminate the effect of the isolated index on the preceding spacing. This operation was repeated until all of the indices had been thoroughly arranged. By virtue of the fact that the effect of the previously isolated indices is eliminated in each spacing, the S(i) and R(i) functions are not monotonically decreasing functions.

It is apparent from Table 1 that the absorption bands at 470-480, 1380-1390, 1180-1190, 1260-1270, 570-580, and 660-670 cm^{-1} bear the greatest information regarding the affiliation of the sample with the cis or trans series. The absorption bands at 1000-1170 cm^{-1}, which, however, according to the data in [7], are characteristic for the 1,3-dioxane ring, are of somewhat lesser informational value. The expression for the determination of R(i) isolates the absorption bands at 1100-1110, 585-595, 1120-1130, 1150-1160, and 1380-1390 cm^{-1} as the most informative bands.

The methods of the theory of graphs, in particular, the algorithm for the isolation of the maximum full and empty subgraphs by means of a Minsk-32 computer, were used to construct the groups of wave numbers [8]. A preliminary analysis of the data in Table 2 indicates that the starting set of frequencies practically does not contain groups of intimately correlated indices.

Those indices for which $\rho_{ij} \geq 0.5$ were considerd as intimately correlated indices. This threshold was selected arbitrarily to a considerable degree, inasmuch as the groups of indices for which $\rho_{ij} < 0.5$ also can be used as informative groups, but the error in discrimination due to losses of information during

TABLE 1. Results of Arrangement of the Absorption Bands of the IR Spectra in the Order of Decreasing Informational Value for 2,5-Dialkyl-1,3-dioxanes

Frequency range for S(i)	No. of bands	*S(i)*	Frequency range for R(i)	No. of bands	*R(i)*
470—479	50	20,56	1100—1110	23	6,8
660—670	40	11,42	585—595	42	5,2
540—550	45	10,94	1190—1200	15	4,13
1030—1040	26	11,53	1120—1130	21	2,97
1410—1420	4	11,28	480—490	49	2,68
830—840	36	10,48	1240—1250	10	1,89
1120—1130	21	10,47	1140—1150	20	2,08
1225—1240	12	10,65	1470—1480	1	1,95
970—980	30	10,06	1380—1390	6	1,61
1110—1120	22	9,66	1150—1160	19	1,74
1190—1200	15	10,01	1460—1470	2	1,66
640—660	41	9,96	950—960	32	1,28
500—520	48	10,09	660—670	40	1,28
580—590	42	10,53	530—540	46	1,08
1140—1150	20	11,27	1260—1270	9	1,38
1390—1400	5	12,08	1290—1310	8	1,15
570—580	43	13,99	780—790	37	0,8
1380—1390	6	15,55	1160—1170	18	0,91
1450—1460	3	12,46	910—920	34	0,98
1260—1270	9	14,49	1010—1020	28	0,66
1180—1190	16	14,82	550—570	44	0,58
1240—1250	11	13,07	520—530	47	0.51
1210—1220	13	12,20	1370—1380	7	0,55
1040—1050	25	3,55	690—700	38	0,44

TABLE 2. Grouping of the Absorption Bands as a Function of the Coefficient of Pair Correlation between Them

Group No.	Frequency groups			
		Intimately correlated		
1	1385—1400,	1380—1385,	1370—1380	
2	690—700,	670—680,	660—670,	640—660
		Weakly correlated		
1	1470—1480,	1460—1470,	1450—1460,	1420—1440
2	1290—1310,	1260—1270		
3	1250—1260,	1240—1250,	1220—1240,	1210—1220
4	1200—1210,	1190—1200,	1180—1190,	1170—1180
5	1160—1170, 1110—1120	1150—1160,	1140—1150,	1120—1130
6	1100—1140,	1060—1080,	1040—1050	
7	1030—1040,	1020—1030		
8	1010—1020,	990—1010,	970—980	
9	960—970,	940—960,	910—930	
10	860—880,	830—850,	730—830	
11	590—600,	570—590,	550—570	
12	540—550,	530—540,	520—530	
13	500—520,	480—500,	470—480	

removal of some of the indices from the group is considerably larger in this case (Table 3). We note that the correlation matrix was almost devoid of strongly correlated indices. This can be explained by the fact that the correlation was made between individual absorption bands of the correlated wave number rather than between the absorption bands due to a strictly determined form of vibrations of any structural element or bond. The rigorous assignment of the correlated absorption bands in the IR spectra of the investigated substances would naturally lead to better results in discrimination. However, this sort of assignment by calculation of the vibrations is as yet difficult.

Two groups of intimately correlated indices were isolated. The coefficient of the correlation between each pair of wave numbers is greater or equal to 0.7 in the first group and greater or equal to 0.5 in the second. The correlated indices as functions of their elements can be brought into line with each of the groups. For this, one can, for example, use regression analysis.

It is apparent from Table 2 that there are a considerable number of groups, the correlation within which is less than or equal to 0.2. This is not difficult to understand if one takes into account the fact that

TABLE 3. Results of Discrimination of the Configurations of Substituted 1,3-Dioxanes

Algorithms	Error in discrimination of the configuration of compounds for various algorithms (in %)			
	2,4,5-trialkyl-1,3-dioxanes		2,4,5-trialkyl- and 2,5-dialkoxyalkyl-1,3-dioxanes	
	complete set of frequencies	minimal set: 15 frequencies	complete set of frequencies	minimal set: 15 frequencies
Method of potential functions	16.3	17	25	30
Albega	23	31	10	20
Linear programming	33	40	40	50

TABLE 4. Arrangement of the Absorption Regions in Order of Decreasing Informative Character

2,4,5-Trialkyl-1,3-dioxanes		2,5-Alkoxyalkyl-1,3-dioxanes		2,4,5-Trialkyl- and 2,5-dialkyl-5-alkoxyalkyl-1,3-dioxane	
No. of bands	wave number region	No. of bands	wave number region	No. of bands	wave number region
47	510—520	48	500—510	42	600—620
46	520—540	45	540—560	4	1410—1420
40	650—670	42	600—620	14	1200—1210
29	990—1010	41	630—640	44	550—560
50	470—480	20	1140—1150	38	690—700
31	960—970	40	650—670	36	830—840
5	1390—1400	14	1190—1200	32	940—950
48	500—510	37	800—820	19	1150—1160
30	680—690	32	940—950	17	1160—1170
42	600—620	7	1360--1370	11	1240—1250

deformation vibrations of methyl and methylene groups appear at 1370-1400 cm^{-1}, while the differences between the isomers reduce to a change in the orientation of one of the substituents in the 5-position of the 1,3-dioxane ring.

The absorption at 400-700 cm^{-1} is due to the skeletal vibrations, which are apparently also sensitive to a change in the three-dimensional structure of the molecule. The latter is directly associated with the axial or equatorial orientation of the alkyl group in the 1,3-dioxane ring. The wave numbers of the absorption bands in the other regions of the IR spectra are weakly correlated (Table 2). It is characteristic that all of the absorption bands at 1000-1200 cm^{-1}, i.e., the "fingerprint region," which is not suitable enough for the identification of organic substances, are related to them. In addition, the absorption bands observed in this region of the IR spectra are, according to the data in [2, 7], characteristic for 1,3-dioxane systems and are less subject to the effect or are not additive to the effect of a change in the vibrational coordinates of other atoms or groups of atoms of the molecules.

The above reasoning is also in good agreement with the data of the quantitative algorithm for minimization of the descriptions. In this variant for the selection of the most informative wave numbers, the informative character of the wave numbers is evaluated by means of the following expression:

$$I(f_i) = -\lg \frac{\sigma_{i1}^{P_1} \cdot \sigma_{i2}^{P_2}}{\sigma_i}, \qquad (2)$$

where σ_i is the standard deviation of the wave number from the average value in both classes, i.e., in the isomers, simultaneously σ_{i1} and σ_{i2} are the standard deviations of the wave numbers from the average in the cis and trans isomers, respectively, and P_1 and P_2 are the probabilities of the classes. In this case, we have assumed that $P_1 = P_2 = 0.5$. The informative character of the groups of wave numbers can also be evaluated from the following expression:

$$I(f_1, \ldots, f_k) = -\lg \frac{\prod_{i=1}^{k} \sigma_{i1}^{P_1} \cdot \prod_{i=1} \sigma_{i2}^{P_2}}{\prod_{i=1}^{k} \sigma_i} + \lg \frac{|R_1|^{P_1} \cdot |R_2|^{P_2}}{|R|}, \qquad (3)$$

where $|R|$, $|R_1|$, and $|R_2|$ are the determinants of the matrices of correlation of the wave numbers of the cis and trans isomers in the entire set of compounds and for each class separately. The informative character of the wave number was calculated in each spacing of the operation of the algorithm from formula (2), and that wave number for which $I(f_i)$ reaches a maximum was selected. Its effect was eliminated by means of a partial correlation by the method described in [6], and the next wave number was selected in the same manner in the set of remaining wave numbers. This operation was carried out until all of the elements of the description were completely arranged. The results of the treatment of the blocks of starting information are presented in Table 4. As seen from this table, the absorption bands at 570-670, 820-860, 960-980, 1100-1200, and 1280-1380 cm^{-1} are informative in the sense of the configurational assignment of the investigated compounds to the cis or trans series. One is easily convinced that the set of wave numbers presented here, which are the most informative, are in satisfactory agreement with the data in Table 1. However, this method gives too broad intervals of wave numbers (100 cm^{-1}) in which one should seek the spectral indices of the individual stereoisomers. In this connection, it is less suitable in this form for the solution of the problem at hand. In addition, the application of the quantitative algorithm of minimization of the description to the analysis of the IR spectra of stereoisomeric 4,5-substituted 1,3-dioxanes [9] gives narrower absorption regions (see Table 4). This same result was obtained by joint analysis of the IR spectra of 2,5- and 4,5-substituted 1,3-dioxanes.

Having at our disposal data on the most informative character of the individual absorption bands in the IR spectra of cis and trans-substituted 1,3-dioxanes, we set out to ascertain the possibility of the recognition of cis and trans isomers by means of a computer [10]. For this, we had to select a certain algorithm from the relatively large number of algorithms that are used in the recognition of forms by means of a computer. Considering the fact that there are no strictly defined recommendations with respect to the use of one or another algorithm for the solution of a concrete problem, especially for the solution of our problem, we selected the method of potential functions [11], "Albega" [10], and linear programming [12]. Each of these algorithms is characterized by a certain measure of allowance for the topology of the sets.

The instructional sequence consisted of 22 2,5-substituted 1,3-dioxanes, while the examining sequence consisted of 10 such 1,3-dioxanes. The instruction was accomplished both with the complete set and with the set of the most informative wave numbers obtained as a result of minimization. The results of an analysis of the examining sequence are presented in Table 3. It is apparent from Table 3 that the error in recognition of the configurations of the stereoisomeric substituted 1,3-dioxanes depends on the algorithm used. The use of the "Albega" algorithm gives the best results for the various substituted 1,3-dioxanes. This can be explained by the fact that it is precisely this algorithm that best takes into account the topology of the sets of cis and trans isomers.

The results of the examination indicate that the highest accuracy in the configurational assignment of the individual stereoisomers to the cis or trans series by means of a computer is achieved when the complete set of absorption bands is used as the starting data. The error in recognition in this case ranges from 10 to 30%. As seen from Table 3, the recognition of the stereoisomers only with respect to 15 of the most informative wave numbers gives poorer results. At first glance, this contradicts the above-stated concepts regarding the set of the most informative bands, but this is really not so. In fact, first, a complete analysis of a spectrum is always more informative than an analysis of its parts; second, the most informative bands are nonequivalent with respect to their own spectrum (reduction of the number of them may give better results); third, the deviations in the course of the examination with respect to all of the bands and with respect to the most informative bands are small within the limits of the possibilities of this method. The fact that the indicated bands are the most informative ones does not raise any doubts and is additionally confirmed by the fact that the inclusion in the examination sequence of wave numbers observed in a mixture of isomers leads to the assignment by the computer of substances simultaneously to both classes (i.e., to the cis and trans series simultaneously), as should be expected.

LITERATURE CITED

1. A. V. Bogat-skii, Yu. Yu. Samitov, S. A. Petrash, A. I. Gren', M. Bartok, and G. Bozoki-Bartok, Zh. Organ. Khim. (1974, in press).
2. A. I. Gren', in: Problems in Stereochemistry [in Russian], Vol. 2, Izd. Kievskogo Univ. (1972), p. 76.
3. A. V. Bogat-skii, A. I. Gren', Yu. Yu. Samitov, L. M. Krinitskaya, L. N. Vostrova, V. N. Somchinskaya, V. P. Mamontov, and T. I. Davidenko, Khim. Geterotsikl. Soedin., 582 (1971).
4. N. V. Terent'ev, Vestn. LGU, No. 9, 137 (1959).

5. G. N. Vostrov and Yu. V. Rublev, in: Scientific-Technical Information [in Russian], Series 2, No. 4, Izd. VINITI (1973), p. 8.
6. G. Kramer, Methods of Mathematical Statistics [Russian translation], Nauka, Moscow (1948), p. 310.
7. K. Ledwoch, Z. Anal. Chem., 197, 323 (1962).
8. A. A. Zykov, Theory of Finite Graphs [in Russian], Novosibirsk (1969).
9. A. V. Bogat-skii, Yu. Yu. Samitov, A. I. Gren', and S. G. Soboleva, Khim. Geterotsikl. Soedin., 893 (1971).
10. G. N. Vostrov, I. B. Gernega, M. L. Varlamov, and G. A. Manakin, in: Methods and Systems for the Treatment of Experimental Information [in Russian], Kiev (1972), p. 31.
11. M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér, Methods of Potential Functions in the Theory of Computer Instruction [in Russian], Nauka, Moscow (1970).
12. Yu. V. Devingtal', Izv. Akad. Nauk SSSR, Ser. Tekhn. Kibernetika, No. 1, 162 (1968).