Theory of Correlation Tables. 2

G. CSONKA and T. VESZPRÉMI*

Department of Inorganic Chemistry, Budapest Technical University, 1521-H, Budapest, Hungary

Received October 29, 1979

A process has been given for the optimization of spectroscopical correlation tables containing complex primary and subfragments. The method is based on information theory and provides the quantitative evaluation of the usefulness of any change within a given correlation table. In this way, for example, it shows in which case it is expedient to specialize a fragment and also when it is worth drawing together several subfragments into one fragment. The process can easily be expressed as a computer algorithm, and in this way it can be useful for a set of computerized systems for spectrum analysis.

The previous paper[1] headlighted a possible mathematical model of the correlation tables and introduced an optimization process through a simple example. It was demonstrated how much information could be gained or lost if a given chemical group in a molecule is specified according to its environment. According to the model, the individual groups are totally independent and equal in rank. This means that, if a group is specified further, the new arrangement excludes the fact that all of the new fragments derive from a single group. Because of this approach the amount of information to be gained can decrease in comparison to the original condition in the case of a given fragment specification. Obviously this has no appeal to common sense because, even if group A were specified, it is clear that the new fragments are derived from A so information cannot be lost.

Actually the correlation table used by a spectroscopist is of a more complex structure. Within the primary fragments adjoining each other are subfragments also adjoining each other but subordinate to the primary fragments. In fact, increased specificity usually means creating several subfragments within a primary fragment. The basic difficulty encountered in discussing correlation tables of this type from the point of view of information theory is that there is no a priori knowledge at hand about the probability of the presence of the individual fragments. It is evident that in this case the sum of the probabilities of the presence of the subfragments will equal the probability of the presence of the primary group.

But how can this probability be divided among the subfragments? A second question immediately arises: what is the probability of the presence of the individual primary fragments? These questions can be answered only by taking into account the formulation of the correlation table.

The following pages describe through a new example the optimum formation of a correlation table containing structurally complex primary and subfragments.

MATHEMATICAL BASIS OF SETTING UP A CORRELATION TABLE

In order to set up a correlation table, a data bank containing assigned spectra is needed. In the case of a practicing spectroscopist this data bank is furnished by the spectra analyzed through the years. The spectra are given with a finite accuracy; in the case of ^{1}H NMR spectroscopy, it is 0.01 ppm. There is no spectrum given more accurately than that in the data bank and the results of measurements will not be registered more accurately.

Let B be a random variable defined by the set of possible results of the measurements $\{y_1, ..., y_j, ..., y_{max}\}$.

Let the event B_j signify that y_j is the result of the measurement, so $B = y_j$. It is important to note that the measured quantity is continuous so event B_j means that the signal is within the rounding interval related to y_j: $y_j^l < y < y_j^u$ where y_j^l and y_j^u are the lower and upper limits, respectively.

Table I. Structural Elements in ^{1}H NMR Spectroscopy

-OH	-SH	Csp_2H
-COH	-ArH	CspH
$-NH_2$	$-CH_3$	$-\underset{|}{CH}-$
-NH	$-CH_2-$	

After this let us select simple structural elements which are relevant to the spectroscopy examined here. In Table I some structural elements which may be relevant in ^{1}H NMR are gathered.

Let us differentiate the structural elements according to their chemical environment and call the larger structural elements originated in this manner groups. Environments are not given absolute consideration.

Let A be a random variable defined by the set of possible groups $\{x_1, x_2, ..., x_k, ..., x_N\}$. Let event A_k symbolize the selection of kth group, so $A = x_k$. N is the number of differentiated fragments.

Let us select from the data bank the compounds which contain the kth group and let us count the number of times this fragment is present; this will be symbolized by s_k. Let us do this calculation with all the groups. The total number of groups $s_t = \sum_k s_k$. The random variable A takes its kth value with the probability of $P(A_k) = P_k = s_k/s_t$.

In the case of a selected x_k group, the following can be established. Let us take the event when the value of B is smaller than an arbitrary value of y ($B < y$). The probability of this is: $p_k(B < y)$. The function $F_k(y) = p_k(B < y)$ is the distribution function of the random variable B if the condition $A = x_k$ is present. This distribution function is illustrated in Figure 1.

The probability that the jth signal is found in the presence of the kth fragment is:

$$F_k(y_j^u) - F_k(y_j^l) = p(B = y_j|A_k) = p(B_j|A_k) \qquad (1)$$

The probability of the simultaneous occurrence of both events is

$$P_{kj} = p(B_j|A_k)p(A_k) \qquad (2)$$

The distribution function of the signal taking into consideration all groups is:

$$F(y) = p(B < y) = \sum_{k=1}^{n} p_k(B < y)p(A_k) \qquad (3)$$

so the distribution function of the signal is the mean average of the distribution function of the individual groups.

In compiling a correlation table, only the two extreme limits of the signal distribution are noted (y_{kl} and y_{ku} in Figure 1) and within these limits a uniform distribution is assumed. Later it is accepted that the signal of the group always appears within this interval.

0095-2338/80/1620-0239$01.00/0

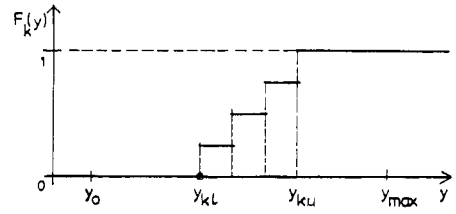

Figure 1. Signal distribution function of a fragment.

Table II. Data and Probabilities in a Fictive Correlation Table

j	y [ppm]	M_{kj}	M_j	$p_{kj}\cdot10^3$	$p_j\cdot10^3$
1	2,1	5	25	1	5
2	2,2	30	45	6	9
3	2,3	10	40	2	8
4	2,4	10	35	2	7
5	2,5	40	60	8	12
6	2,6	5	25	1	5
$\sum$		s_k=100		p_k=2·10^{-2}	4.6·10^{-2}

OPTIMIZATION OF STRUCTURALLY COMPLEX CORRELATION TABLES

Let us solve the following problem.

There are at our disposal $N = 200$ kinds of fragments and a data bank that contains assigned spectra in which each fragment is represented by an average of 25 compounds. This means that the correlation table is made up on the basis of 5000 analyses ($s_t = 5000$). Let us select the kth fragment which appears for a total of $s_k = 100$ times. The relative frequency of the kth fragment is $r_k = 100/5000 = 0.02$; let us take this as the probability of the fragment, so $p_k = 0.02$.

Because of rounding, shifts are given in discrete values. Let us count how many times this fragment gives a signal corresponding to the jth shift, and let us symbolize this number by M_{kj}. Let us count how many fragments give the jth shift apart from k. The sum of the two, M_j, shows the total number of fragments giving the jth signal. The data of Table II show the results of the calculation.

The calculation of probabilities is done in the following way:

$$p_{kj} = M_{kj}/s_t; \qquad p_j = M_j/s_t \tag{4}$$

and, of course,

$$p_k = \sum_j p_{kj}$$

In the table let y signify the measured characteristic of the spectrum (e.g., at ^{1}H NMR the chemical shift is in ppm; the rounding interval in this example is 0.1 ppm wide).

Let us divide the kth fragment in two possible ways. On the basis of information theory let us decide if it was worth making the division.

In case I fragments k_1' and k_2', are created; in case II, fragments k_1'' and k_2''. Table III shows how many compounds give shifts in the new sections.

The calculation of the probabilities present in the table is done on the basis of the following relationship:

$$p_{kj} = M_{kij}/s_t \qquad (i = 1, 2) \tag{5}$$

The following relationship is true for all j's:

$$M_{kj} = M_{k1j} + M_{k2j} \tag{6}$$

Let us calculate in both cases the change of information on the basis of the following equation:

$$\Delta I = I' - I \tag{7}$$

Table III. Signals and Probabilities in Two Different Divisions

		I				II			
j	y	M_{k_1j}	$p_{k_1j}\cdot10^3$	M_{k_2j}	$p_{k_2j}\cdot10^3$	M_{k_1j}	$p_{k_1j}\cdot10^3$	M_{k_2j}	$p_{k_2j}\cdot10^3$
1	2.1	2	0.4	3	0.6	0	0	5	1
2	2.2	15	3	15	3	0	0	30	6
3	2.3	5	1	5	1	0	0	10	2
4	2.4	5	1	5	1	10	2	0	0
5	2.5	20	4	20	4	40	8	0	0
6	2.6	3	0.6	2	0.4	5	1	0	0
$\sum$		50	10^{-2}	50	10^{-2}	55	1.1·10^{-2}	45	0.9·10^{-2}

where I' is the information after the specialization and I is the initial information.

The equation for the calculation of the information can be found in the previous paper.[1] The information is additively composed of the contributions of individual fragments. The change of information in this case results from the change of the kth fragment:

$$\Delta I = \sum_{l=k_1,k_2} \sum_{j=1}^{6} p_{lj} \log_2 \frac{p_{lj}}{p_l p_j} - \sum_{j=1}^{6} p_{kj} \log_2 \frac{p_{kj}}{p_k p_j} \tag{8}$$

By substituting the probabilities present in Table III into eq 8, we get:

$$\Delta I' = 9.249 \times 10^{-2} - 9.243 \times 10^{-2} = 6 \times 10^{-5} \tag{9}$$

In case II we can proceed in the same way, and the result is

$$\Delta I'' = 11.228 \times 10^{-2} - 9.243 \times 10^{-2} = 1.985 \times 10^{-2} \text{ (bit)} \tag{10}$$

It is evident that the fragment classification realized in case II provides substantially more information.

Let us analyze, in both cases, the relationships among the changes of the initial entropy $H(A)$, the information $I(A,B)$, and the conditional entropy ($H_B(A)$).

The change of $H(A)$ in case I is:

$$\Delta H'(A) = -p_{k_1} \log_2 p_{k_1} - p_{k_2} \log_2 p_{k_2} + p_k \log_2 p_k \tag{11}$$

because

$$p_{k_1} = p_{k_2} = p_k/2$$

$$\Delta H'(A) = p_k \log_2 p_k/p_{k_1} = p_k \log_2 2 = p_k = 0.02 \text{ (bit)} \tag{12}$$

$$\Delta H_B'(A) = \Delta H'(A) - \Delta I' = 0.02 - 6 \times 10^{-5} \simeq 0.02 \tag{13}$$

In case II

$$\Delta H''(A) = -9 \times 10^{-3} \log_2 (9 \times 10^{-3}) - 1.1 \times 10^{-2} \log_2 (1.1 \times 10^{-2}) + 0.02 \log_2 0.02 \tag{14}$$

$$\Delta H''(A) = 1.985 \times 10^{-2} \text{ (bit)}$$

$$\Delta H_B''(A) = \Delta H''(A) - \Delta I'' = 0 \text{ (bit)} \tag{15}$$

Figure 2 gives an example.

On the basis of Figure 2, it is clear that in case I the conditional entropy ($H_B'(A)$) increased at the same rate as initial entropy; therefore, the information gain is practically zero. In case II the increase of the initial entropy causes all the increase of the information; the conditional entropy is unchanged. Thus it can be established that it is expedient to make the change that infers the greater proportion of

$$\Delta I(A,B)/\Delta H(A) = R \qquad (\text{where } 0 \le R \le 1)$$

The inverse operation can also be done. If we reduce case II to its starting condition, the decrease of the initial entropy

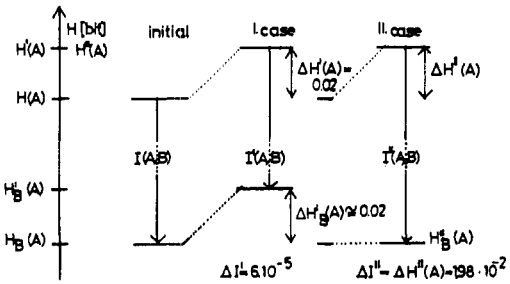

Figure 2. Entropy level change in two different cases.

(now $H''(A)$) causes only the decrease of the information without decreasing the conditional entropy. It is not expedient to execute this change. On the other hand, in case I the conditional entropy decreases together with the initial entropy (now $H'(A)$) while information remains the same. Obviously it is expedient to make such a change because with the unchanged information the size of the correlation table decreases.

The correlation tables contain the probabilities in a simplified form because one assumes a uniform distribution. Taking this into consideration, information theory can be used in an unchanged form.

In this case

$$p_{kj} = 0.02/6 = 3.33 \times 10^{-3}$$

is constant for all j's.

In case I

$$p_{k_1 j} = p_{k_2 j} = p_{kj}/2, \qquad p_{k_1} = p_{k_2} = p_k/2$$

After the substitution of $p_{k_1 j}$, $p_{k_2 j}$, p_{k_1}, and p_{k_2} according to their values and having done the reduction, the change of the information on the basis of eq 8 is:

$$\Delta I' = \sum_j \frac{p_{kj}}{2} \log_2 \frac{2p_{kj}}{2p_k p_j} - \sum_j p_{kj} \log_2 \frac{p_{kj}}{p_k p_j} = 0 \quad (16)$$

In case II the following conditions are satisfied:

$$p_{k_1} = \frac{p_k}{1.818}; \qquad p_{k_1 j} = p_{kj}, \text{ if } j = 4, 5, 6, \text{ otherwise } 0$$

$$p_{k_2} = \frac{p_k}{2.222}; \qquad p_{k_2 j} = p_{kj}, \text{ if } j = 1, 2, 3, \text{ otherwise } 0 \quad (17)$$

The substitution of relations 17 into eq 8 gives:

$$\Delta I'' = \sum_{j=1}^{3} p_{kj} \log_2 \frac{2.222 p_{kj}}{p_k p_j} + \sum_{j=4}^{6} p_{kj} \log_2 \frac{1.818 p_{kj}}{p_k p_j} - \sum_{j=1}^{6} p_{kj} \log_2 \frac{p_{kj}}{p_k p_j} \quad (18)$$

After having done the reduction we get:

$$\Delta I'' = \sum_{j=1}^{3} p_{kj} \log_2 2.222 + \sum_{j=4}^{6} p_{kj} \log_2 1.818 =$$

$$p_{k_2} \log_2 \frac{p_k}{p_{k_2}} + p_{k_1} \log_2 \frac{p_k}{p_{k_1}} = 1.985 \text{ (bit)} \quad (19)$$

By comparing eq 16 and 19 with eq 9 and 10, it is evident that the results are similar.

The change of the initial entropy can be calculated on the basis of eq 12. On the basis of the detailed observation of the mathematical equations, the difference between the information calculated in two different ways is that the amount of information calculated on the basis of the original distribution function is more precise.

Based on the relationships obtained so far, the following can be established. The maximal gain of information that can be obtained by the increased specificity in fragments chosen corresponds to the increase of the initial entropy. This occurs when the intervals of the fragments created during a specialization do not overlap (see case II).

The increase of the initial entropy is highest when the specialization is made into fragments of equal probability. (These characteristics can be understood easily on the basis of the relationships.)

From a practical viewpoint it is expedient to take into account the increasing memory requirements. Emphasized consideration of the environment goes with the increase of the number of fragments and also with the increase of memory need. The two variables can be compared based on the following relation:

$$\Delta I/I > \Delta M/M \quad (20)$$

where $\Delta I/I$ is the relative increase of information and $\Delta M/M$ is the relative increase of the memory requirements.

With a reduction of the number of the fragments, the relation is the opposite because of the negative sign. As the memory is in general a limited value, we might need the optimization of the correlation table with a constant memory. In this case all changes which cause an increase of information are useful.

The example above gives an optimal algorithm for the setup of a correlation table if it is constructed on the basis of an assigned collection of spectra. However, at this point, because of the finite spectra collection, it is also the result of the specification that the observed fragment is present in fewer and fewer compounds. That is why the conclusions are drawn on the basis of fewer observations, so the results are more uncertain.

REFERENCES AND NOTES

(1) Veszprémi, T. J.; Csonka, G. I., *J. Chem. Inf. Comput. Sci.*, preceding paper in this issue.