Journal of Chromatography, 122 (1976) 535–551

CHROM. 9074

THE SELECTION OF REPRESENTATIVE SUBSTANCES BY OPERATIONS RESEARCH TECHNIQUES

APPLICATION TO THE SELECTION OF FUNCTIONAL PROBES FOR GAS–LIQUID CHROMATOGRAPHY

H. DE CLERCQ

Farmaceutisch Instituut, Vrije Universiteit Brussel, Paardenstraat 67, B-1640 Sint-Genesius-Rode (Belgium)

M. DESPONTIN and L. KAUFMAN

Centrum voor Statistiek en Operationeel Onderzoek, Vrije Universiteit Brussel, Pleinlaan 2, B-1040 Brussels (Belgium)

and

D. L. MASSART

Farmaceutisch Instituut, Vrije Universiteit Brussel, Paardenstraat 67, B-1640 Sint-Genesius-Rode (Belgium)

SUMMARY

In gas–liquid chromatography, schemes have been developed for the characterization of stationary phases based on the measurement of the retention indices of a number of functional probes. A problem that arises in this context is how many probes should be used and which solutes should be chosen to function as a probe. A completely objective, *i.e.* mathematical, approach to the selection of probes is proposed.

By expressing the similarities between each pair of solutes as a distance, a complete, non-directed graph can be constructed in which the solutes are represented by nodes, linked together by edges, the values of which are given by the distance. The problem of finding representative solutes is then reduced to selecting the nodes for which the sum of the values of the edges between the unselected nodes and the nearest selected node is minimal. Two solutions are possible: a heuristic solution, which gives an approximation of the optimal choice of probe, and a completely optimal solution in which the integer programming problem is solved by using a branch and bound method.

The work was carried out on two sets of retention indices, namely those given by Rohrschneider (23 columns × 30 substances) and McReynolds (25 columns × 68 substances). Sets of p probes were selected using the described methods for $p = 1, 2, \ldots 20$ for both sets of retention indices. The results for $p = 3, 4, 5$ were compared with those proposed in the literature and it was found that the probes selected by us do allow a better prediction of the retention indices of the other substances in the sets of retention indices than those proposed by other workers.

INTRODUCTION

In a recent report[1], it was shown that operations research (O.R.) techniques are useful in the solution of some analytical problems. It is the purpose of this paper to try to prove this further by demonstrating how to select representative substances by O.R. A typical example of the use of representative substances is the characterization of gas–liquid chromatographic (GLC) phases with functional "probes". In analytical separation chemistry, it has often been found necessary to have methods for the characterization of separation systems. In GLC, for example, schemes have been developed for the characterization of stationary phases, the most widely used of which are those proposed by Rohrschneider[2] and McReynolds[3]. These schemes are based on the measurement of the retention indices of a number of standard solutes or "probes" and they have been used for more formal classification systems using nearest neighbour[4], pattern cognition[5] and numerical taxonomy[6] techniques. Analogous problems are encountered in other branches of separation science. One can refer, for example, to recent work by Rohrschneider[7] and Snyder[8] on solvent classification and Massart and De Clercq[9] on the classification of thin-layer chromatographic systems by numerical taxonomy. All of these classifications are based on similarities in the behaviour of a number of "probes", the retention indices (or R_F values or solubilities etc.) of which are measured in the systems to be classified.

A problem that arises in this context is how many probes should be used and which solutes should be chosen as a probe. It is clear that, the smaller their number, the easier it is to carry out the necessary measurements for the classification of a given separation system. This question has recently received much attention in GLC and has been discussed, for example, by Hartkopf and co-workers[10,11] and Lowry *et al.*[12]. The problem is in fact the following: how can one choose a number of standard solutes so that they are as representative as possible of a given set?

Rohrschneider[2] selected for chemical reasons from a restricted set of 30 and Hartkopf *et al.*[10,11] and Lowry *et al.*[12] from a set of 68 solutes a small number of probes which they thought should be representative of a given interaction between the stationary phase and the chromatographed solute. For example, Hartkopf *et al.*[11] propose the use of benzene (dispersion forces), nitroethane (dipole orientation), *n*-propanol or chloroform (proton donor) and dioxane (proton acceptor); such a selection procedure is necessarily subjective. It was thought that a completely objective, *i.e.* mathematical, approach to the selection of probes would therefore be of interest.

It seems that the selection of representative substances is not an uncommon problem in analytical chemistry. Such a problem has been stated recently by Haken[13], who considered that the usual McReynolds or Rohrschneider constants are not ideally suited for characterizing stationary phases for lipid analysis and that a set of separation factors might be a better solution. One might also suggest the choice of a number of representative lipids. In this paper, we shall confine the discussion to the selection of generally representative probes for GLC.

THE MODEL

The solution proposed is the following. If, in a manner that will be specified

later, the similarities between each pair of solutes can be expressed as a distance, then a complete, non-directed graph can be constructed, in which the solutes are represented by nodes, linked together by edges the values of which are given by the distance. The problem of finding representative solutes is then reduced to selecting the nodes for which the sum of the values of the edges between the unselected nodes and the nearest selected node is minimal. This is a typical location problem, which resembles classical problems in mathematical programming such as the location of warehouses or other service centres. The general problem can be stated as follows: "for a finite number of users, whose demands for a given service are known and must be fulfilled and a finite set of possible locations where a given number p of service centres may be located, select the locations of the service centres in order to minimize the sum of transportation costs of the users".

Mathematically, the problem can be described as follows:

Minimize

$$\sum_i \sum_j d_{ij} x_{ij} \tag{1}$$

subject to

$$\sum_i x_{ij} = 1 \tag{2}$$

$$x_{ij} \leqslant y_i \tag{3}$$

$$\sum_i y_i = p \tag{4}$$

$$y_i \in \{0,1\} \tag{5}$$

$$x_{ij} \in \{0,1\} \tag{6}$$

where

$i = 1, \ldots, n$ and $j = 1, \ldots, n$;
p = number of probes;
d_{ij} = distance between substance j and probe i;
x_{ij} = a variable that determines which probe is representative of substance j; $x_{ij} = 1$ if j is closest to probe i and is therefore represented by i and $x_{ij} = 0$ when it is not the case;
y_i = a variable that determines whether a substance is selected as a probe; $y_i = 1$ when this is the case and $y_i = 0$ when it is not.

Two solutions are possible: a heuristic solution, which gives an approximation of the optimal probe choice, and a completely optimal solution.

The latter cannot be obtained by the more usual linear programming method because the variables can take only the values 0 or 1, and therefore a branch and bound method is used. These methods have been discussed in many textbooks on operational research and linear programming and related methods[14].

Although the method was first proposed in 1954 by Land and Doig[5] no applications in analytical chemistry are known to us and it seems necessary to explain the principle of the method. Branch and bound methods are partial enumeration methods, which consist in partly enumerating the set of solutions in such a manner

that subsets of solutions, which, by using one or other decision criterion (bound), can be shown not to contain the optimal solution, can then be left out. In this way, the set of possible solutions is divided by way of a dichotomic decision tree (branching) in sets that become smaller and smaller until only the optimal node is left.

THEORETICAL

A variable y_i is associated with each element i. The value of y_i is unity if element i is selected, otherwise it is zero. A solution of the problem is defined as the selection of p elements. This corresponds to giving the value unity to p components of vector y and the value zero to the remaining ones. As the number of solutions is finite, it is possible to enumerate all of them and for a given criterion to select the optimal solution. However, as the number of solutions is very large, such an explicit enumeration is prohibitive even for powerful computers.

The branch and bound procedure which will be used proceeds by an implicit enumeration of all solutions. The set of all solutions is separated into small subsets of solutions, using a separation principle, and by examining such subsets the solutions are not all considered separately. During the examination, a subset S is defined as fathomable if one of the following conditions is satisfied:

(1) S does not contain a solution better than the best solution found so far. S can then be eliminated from further consideration.

(2) The best solution of S is better than the best solution found so far. In this case, this solution replaces the old one.

When a subset S is fathomed it must not be separated. The next subset can then be examined.

At the beginning of the branch and bound algorithm, all variables are "free". This means that it has not yet been decided which value the variables will take. The separation principle consists in selecting a variable y_i and constructing two subsets of solutions. The first subset contains all solutions for which $y_i = 1$ and the second contains all solutions for which $y_i = 0$. The first subset to be examined is always the one for which $y_i = 1$. If a subset is examined and fathomable, the method takes the last variable y_i which was set equal to unity and takes the other subset (for which $y_i = 0$). To make it possible to fathom a subset S, a value $B(S)$ is computed, which is a lower bound on the value of the objective function for all solutions belonging to S. If this lower bound is greater than the value of the best solution obtained so far, then S is fathomed.

To construct the objective function of the problem, it is necessary to define variables x_{ij} which are equal to unity if the element i is the selected element nearest to j and equal to zero otherwise. The objective function is then given by eqn. 1 and the constraints are given by eqns. 2–6. Constraints 2 express that for an element j only one distance d_{ij} must be taken into account in the objective function. Constraints 3 express that if an element i is not selected and therefore $y_i = 0$, then the variables x_{ij} are all zero and the distances between i and all other elements are not considered. Constraint 4 expresses that p elements must be selected.

This problem is solved by a branch and bound algorithm. As a subset is reached after some of the variables y_i have been fixed, the following definitions can be used to characterize this subset:

K_0 = set of indices of variables set equal to 0
$= \{i \mid y_i = 0\}$
K_1 = set of indices of variables set equal to 1
$= \{i \mid y_i = 1\}$
K_2 = set of indices of free variables
$= \{i \mid i \notin K_0 \cup K_1)\}$

The cardinals of these sets are denoted by m_0, m_1 and m_2, respectively. Using these definitions, it is possible to restate problem 1–6 as it occurs with this subset:
Minimize

$$\sum_{i \in K_1 \cup K_2} \sum_j d_{ij} x_{ij} \tag{7}$$

subject to the constraints

$$\sum_{i \in K_1 \cup K_2} x_{ij} = 1 \tag{8}$$

$$x_{ij} \leqslant y_i \qquad i \in K_2 \tag{9}$$

$$m_1 + \sum_{i \in K_2} y_i = p \tag{10}$$

$$y_i = 1 \qquad i \in K_1 \tag{11}$$

$$y_i = 0 \qquad i \in K_0 \tag{12}$$

$$y_i \in \{0,1\} \qquad i \in K_2 \tag{13}$$

$$x_{ij} = 0 \qquad i \in K_0 \tag{14}$$

$$x_{ij} \in \{0,1\} \qquad i \in K_1 \cup K_2 \tag{15}$$

It is impossible to explain in detail here how the problem stated in this way is solved; the complete solution method is given in ref. 16.

DATA SETS AND CHOICE OF DISTANCES

Most of the work was carried out on two sets of retention indices, namely those given by Rohrschneider[2] and McReynolds[17]. The solute composition of the latter was given by Hartkopf[10]. It should be also noted that McReynolds' work included all solutes used by Rohrschneider. This does not necessarily mean that the results obtained by using McReynolds' data set are better than those obtained by using Rohrschneider's data, and in fact it would seem instead that McReynolds' set could introduce some bias because it contains a disproportionate number of substances with the same functional group. For example, there are 13 aliphatic saturated alcohols (from 68 solutes), while there are only 4 such solutes (from 30) in Rohrschneider's set. On the other hand, the variety of substances included in McReynolds' set is larger. This discussion of data sets leads to the first of the two main objections which, according to us, can be made against the present approach. The purpose of

the selection is to obtain probes that are as representative as possible of all solutes which could be chromatographed. Therefore, in theory, the selection procedure should be based on the universe of all GLC data. In practice, of course, this is not possible and necessarily a restricted set must be chosen. This approach already contains some subjectivity and may lead to erroneous conclusions when a biased set is used. It seems very probable, for example, that if 20 ketosteroids were added to one of the sets discussed above, that at least one of them would have been selected as a probe. Also, when a small group (or even a pair) of closely related substances that are very different from the other solutes are included, it is probable that one of these would be selected. In fact, this occurs here in the selection of the probe 1,1-difluorotetrachloroethane, which is discussed later. It is also possible that a single substance that is very different from all others would be selected.

There is, therefore, no doubt that this objection is valid. It should be realized, however, that this is also true for all of the other work which has been published on the subject of the use of probes in GLC. Rohrschneider, whose indices seem to be generally accepted at present, chose his probes using the data mentioned above and it should be also noted that much thought was given to the composition of the set from which they were chosen. Rohrschneider states: "We chose 30 compounds, from 25 different series, containing the atoms C, H, O, S, N, F, Cl, Br and I and 12 different functional groups".

In this study, the distance $(1-\varrho)$ was used, where ϱ is the linear correlation coefficient. This constitutes the second objection which can be made against the present selection procedure. The correlation coefficient cannot be used in statistical tests of significance in this instance because the distribution of the retention indices of a solute over the 25 columns is not normal, as can be observed when carrying out a χ^2 test. Even then, ϱ can be considered as a measure of similarity in chromatographic behaviour. A second, more severe, difficulty when using ϱ is that, for the alkanes 2,4-dimethylpentane, 3-methylheptane, 2-methylheptane, 2,2,3-trimethylhexane and cyclohexane, the range of retention indices is so small that the values of their correlation coefficients have no significance because they are determined by the random error in the retention indices and are therefore artificially low. This forced us to eliminate these substances from the data sets. This was also necessary for 2-iodobutane because it was considered[10] that some of the retention indices given for this solute must be erroneous. Consequently, only 62 of McReynolds' 68 substances will be used in the present study.

PREDICTION OF RETENTION INDICES BY MULTIPLE REGRESSION

The purpose of this selection procedure is to select better functional probes, which means that they allow a better prediction of retention indices of other substances. Such a prediction can be carried out by using a model resembling that used by Rohrschneider, based on linear regression. As there has been controversy about the correct mathematical way of carrying out this prediction (see, for example, ref. 18), it seems preferable to state in sufficient detail which method was used here.

Using the method of Leary *et al.*[18] for finding the best values of the coefficients for a given phase by least squares, the model can be stated as

$$y_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_p x_{jp} + e_j$$

where $j = 1, 2, \ldots, n$. In matrix notation:

$$y = X\beta + \varepsilon$$

where

- y is the n-element column vector of experimental values of ΔI;
- n is the number of liquid phases;
- X is an observed $n \times p$ matrix;
- p is the number of terms which are calculated from the differences in Kováts retention indices (*i.e.* the number of probes);
- β is a column vector of p unknown parameters;
- ε is an n-element vector of stochastic disturbances due to non-observed and unobservable variables, and mis-specifications of the model.

Making the assumptions

$$E(y) = X\beta$$

and

$$\Sigma_y = \sigma^2 I$$

where Σ = variance–covariance matrix and fitting by ordinary least squares, we obtain the well known result

$$\hat{\beta} = (X'X)^{-1} X'y$$

where $\hat{\beta}$ is an unbiased and minimum variance estimator of β. An unbiased estimator of the variance σ^2 is given by

$$s^2 = \frac{e'e}{n-p} = \frac{y'My}{n-p}$$

where $M = I - X(X'X)^{-1} X'$ and $e = y - X\hat{\beta}$. It can easily be proved that $X'e = 0$. Rohrschneider's assumption that $(y - X\hat{\beta})'i = e'i = 0$, where i is the n-element column vector $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, is generally true only when the multiple regression model contains a constant. Because there is no chemical reason to use such a constant, we reject this assumption.

Using the coefficients determined in this way, the retention indices of all of the substances on all of the stationary phases were calculated for both data sets (Rohrschneider and McReynolds). The calculated values were then compared with the actual values and the standard deviation (σ in Tables IV-VIII) of the errors was calculated. This was thought to be a more significant way of describing the value of the prediction than the more usual calculation of the mean prediction error[2,10,11].

RESULTS

It was found that in many instances the heuristic method yields the optimal solution directly. This was found by systematically comparing the results obtained by the heuristic method and the branch and bound method on a set of 35 substances

TABLE I

PROBES SELECTED FROM McREYNOLDS'[17] SET

Probe	*p*																				
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
1-Octene															×	×	×	×	×	×
2-Octene																				
2-Ethylhexene																				
1-Octyne																			×	×
2-Octyne												×	×	×	×	×	×	×	×	×
Benzene																			×	×
Toluene															×	×	×	×		
Ethylbenzene						×		×		×	×	×	×	×						
Mesitylene									×											
Hydrindane								×	×	×	×	×	×	×	×	×	×	×	×	×
Decalin					×	×	×													
Styrene					×		×													
Ethynylbenzene		×				×	×				×	×						×		
Methanol																				
Ethanol																				
Propanol								×	×							×	×	×	×	×
2-Propanol			×	×		×	×													
n-Butanol										×	×	×	×	×	×					
tert.-Butanol																				
1-Hexanol																				
2-Hexanol																×	×	×	×	×
3-Hexanol					×															
2-Methyl-2-pentanol								×	×	×	×	×	×	×	×	×	×	×	×	×
3-Methyl-3-pentanol																				
1-Hexyl-2-heptanol																				
3-Methyl-3-heptanol																				
Cyclopentanol																				
Allyl alcohol																				
Propionaldehyde		×		×	×	×														
Hexanal																	×	×	×	
Crotonaldehyde			×																	×
Acetone																				

Methyl ethyl ketone					×	×	×	×	×	×	×	×	×	×					
2-Pentanone	×																		
2-Hexanone																			
3-Hexanone																			
n-Butyl acetate	×	×	×	×															
Propyl butyrate																			
Methyl iodide	×	×	×																
1-Iodobutane																			
Ethyl bromide	×																		
Bromopentane		×	×	×	×	×	×	×	×	×									
Chlorohexane																			
Chloroform	×	×	×	×	×	×	×	×			×								
Carbon tetrachloride	×	×																	
1,1,2,2-Tetrachloroethane																			
Fluorooctane																			
1,1-Difluorotetrachloroethane	×	×	×	×	×	×	×	×	×	×	×	×	×				×		
1,2-Difluorotetrachloroethane														×	×	×			
Di-*n*-butyl ether	×	×	×	×	×	×	×												
Anisole	×	×	×	×	×	×	×	×	×	×									
1,4-Dioxane								×	×	×	×	×							
Hexanethiol																			
Propyl sulphide																			
Thiophene				×	×	×	×	×			×	×	×				×	×	
Nitromethane																			
Nitroethane	×	×	×	×	×	×	×	×	×	×	×	×	×	×					
Nitropropane																			
2-Methyl-2-nitropropane																			
Acetonitrile		×																	
Pyridine	×		×	×	×	×	×												×

obtained by simplification of McReynolds'[17] set of 62 substances by 25 phases. This simplification was carried out by calculating the correlation coefficients obtained with the retention indices on the 25 phases for all pairs of solutes and carrying out a numerical taxonomic classification of the set. From each pair of closely related substances, one was deleted. It was found for $p = 2, 3, 4, 5$ that the heuristic solution is the same as the branch and bound solution. The same comparison was carried out on $p = 3, 4, 5, 6$ for the complete set. In this instance both solutions are the same only for $p = 4$, and in the other instances the branch and bound solutions resulted in small decreases in the objective function. As the time necessary to carry out the branch and bound method rapidly becomes prohibitive as p increases, only the heuristic solution was computed for values $p > 7$ for McReynolds' set, while for Rohrschneider's set only heuristic calculations were carried out. The $p = 7$ solution for McReynolds' set is a solution obtained by the branch and bound method, which, however, was not carried out until the end. However, it differs from and is (slightly) better than the heuristic solution.

DISCUSSION

The probes selected from McReynolds' set are given in Table I and those obtained from Rohrschneider's set are given in Table II. Using Rohrschneider's data set, one comes to the conclusion, for example, that if five probes were needed, one should have selected ethanol, propionaldehyde, acetonitrile, dioxane and thiophene. The substances for which each probe is representative (*i.e.* is situated nearest to in the location model) are given in Table III. Rohrschneider[2] proposed ethanol, methyl ethyl ketone, nitromethane, pyridine and benzene.

It can be observed that ethanol is found both among our probes and Rohrschneider's, while each of the other of Rohrschneider's probes is found (see Table III) in a different group represented by one of our probes. This means that the same five factors are represented in both sets of probes. There is, in fact, very little difference between methyl ethyl ketone and propionaldehyde ($\varrho = 0.9995$), acetonitrile and nitromethane ($\varrho = 0.9988$) and benzene and thiophene ($\varrho = 0.9989$). The difference between dioxane and pyridine ($\varrho = 0.9973$) is somewhat greater. However, pyridine is different from all of the other solutes and the one which resembles it most is dioxane. Therefore, one can conclude that Rohrschneider's selection and ours are really analogous.

One can therefore conclude that the selection procedure yields acceptable and logical results. To investigate whether the probes selected are as good as or better than the existing probe sets, a large number of probe sets were used to carry out the prediction procedure described earlier. The results obtained with the now usual number of five probes are given in order of predictive ability for McReynolds' and Rohrschneider's sets in Table IV. In McReynolds' set, probe set 2 is McReynolds' original proposal and probe set 5 the one obtained from Table I by us. In Rohrschneider's set, probe set 1 is Rohrschneider's proposal while set 2 is the one extracted from Table II. In both instances, the results obtained with our selection procedure are acceptable, but still worse than those proposed by McReynolds and Rohrschneider.

We have investigated the reason for this discrepancy in more detail for McReynolds' set. In set 5, the presence of the first and the last probes seems surprising.

TABLE II

PROBES SELECTED FROM ROHRSCHNEIDER'S[2] SET

Probe	*p*											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Benzene			×			×				×	×	×
Ethanol				×	×	×	×	×	×	×	×	×
Methyl ethyl ketone												
Nitromethane												
Pyridine												
2-Ethylhexene												
Toluene												
Styrene									×			
Phenylacetylene						×			×			×
Acetone												
Propionaldehyde					×	×	×	×	×	×	×	×
Crotonaldehyde			×	×								
n-Butyl acetate												
Acetonitrile					×	×	×	×	×	×	×	×
Nitroethane												
Dioxane		×		×	×		×	×	×	×	×	×
Di-*n*-butyl ether						×	×	×	×	×	×	×
Thiophene				×	×		×	×		×	×	×
Chloroform								×		×	×	×
Carbon tetrachloride									×	×	×	×
Methyl iodide												
Ethyl bromide	×											
Difluorotetrachloroethane							×	×	×	×	×	×
n-Propanol												
2-Propanol												
Allyl alcohol												
tert.-Butanol											×	×
Cyclopentanol		×	×									

It is found that decalin is representative only of itself, while 1,2-difluorotetrachloroethane is representative of itself and the other fluorochloroethane compound in the set. As discussed above, this is due to the fact that they are very dissimilar from the other probes. If a substance is selected as a probe but proves representative only of itself or one or two others, it can be considered as unrepresentative of all of the other substances (*i.e.* nearly the complete set). Therefore, it was reasoned that a good solution could have been obtained by selecting as probes only those which are found to be representative of more than two substances. In this instance, one uses the $p = 7$ solution after eliminating decalin and 1,2-difluorotetrachloroethane. The resulting set of probes consists of styrene, ethynylbenzene, 2-propanol, methyl ethyl ketone and nitroethane (probe set 1), which is found to be better than the original proposal by McReynolds.

There is no general agreement that five probes should be used. For example, McReynolds first used ten functional probes and the five listed above were selected by him from this original set of ten. One supplier of GLC phases still uses seven of these probes[19]: benzene, *n*-butanol, 2-pentanone, nitropropane, pyridine, 2-methyl-2-pentanol and 2-octyne (probe set 3 in Table V). Table V also gives the results ob-

TABLE III

SUBSTANCES OF WHICH EACH PROBE IS REPRESENTATIVE IN ROHRSCHNEIDER'S SET

Probes originally proposed by Rohrschneider are italicized.

p	*Substances*	
5	*Ethanol*:	Chloroform, *n*-propanol, 2-propanol, allyl alcohol, *tert.*-butanol, cyclopentanol
	Propionaldehyde:	*methyl ethyl ketone*, acetone, crotonaldehyde, *n*-butyl acetate
	Acetonitrile:	*nitromethane*, nitroethane
	Dioxane:	*pyridine*, 2-ethylhexane, di-*n*-butyl ether, ethyl bromide
	Thiophene:	*benzene*, toluene, styrene, phenylacetylene, carbon tetrachloride, methyl iodide, difluorotetrachloroethane
4	*Ethanol*:	chloroform, *n*-propanol, 2-propanol, allyl alcohol, *tert.*-butanol, cyclopentanol
	Crotonaldehyde:	propionaldehyde, *methyl ethyl ketone*, acetone, *n*-butyl acetate, acetonitrile, *nitromethane*, nitroethane
	Dioxane:	*pyridine*, 2-ethylhexene, di-*n*-butyl ether, ethyl bromide
	Thiophene:	toluene, *benzene*, styrene, phenylacetylene, carbon tetrachloride, methyl iodide, difluorotetrachloroethane
3	Cyclopentanol:	*ethanol*, *n*-propanol, 2-propanol, allyl alcohol, *tert.*-butanol, chloroform, phenylacetylene, difluorotetrachloroethane
	Benzene:	*pyridine*, 2-ethylhexene-1, styrene, toluene, dioxane, di-*n*-butyl ether, thiophene, carbon tetrachloride, methyl iodide, ethyl bromide
	Crotonaldehyde:	*methyl ethyl ketone*, *nitromethane*, acetone, propionaldehyde, *n*-butyl acetate, acetonitrile, nitroethane

TABLE IV

PREDICTION WITH FIVE PROBES

Set	*Rank number*	*Probes*	σ
McReynolds	1	Styrene, ethynylbenzene, 2-propanol, methyl ethyl ketone, nitroethane	9.08
	2	Benzene, *n*-butanol, 2-pentanone, nitropropane, pyridine	9.22
	3	Ethanol, propionaldehyde, acetonitrile, 1,4-dioxane, thiophene	9.63
	4	Benzene, ethanol, methyl ethyl ketone, nitromethane, pyridine	10.42
	5	Decalin, styrene, 3-hexanol, propionaldehyde, 1,2-difluorotetrachloroethane	10.54
Rohrschneider	1	Benzene, ethanol, methyl ethyl ketone, nitromethane, pyridine	5.82
	2	Ethanol, propionaldehyde, acetonitrile, dioxane, thiophene	6.43

tained when using the seven probes obtained from Table I (set 1) and those obtained from Table I using nine probes after elimination of two probes representative of less than three substances (set 2), and two sets of six probes extracted from Table I. Here again, the probe sets proposed by us yield better results than those given in the literature.

TABLE V

PREDICTION IN McREYNOLDS' SET WITH SIX AND SEVEN PROBES

Rank number	*Probes*	σ
1	Decalin, styrene, ethynylbenzene, 2-propanol, methyl ethyl ketone, 1,2-difluorotetrachloroethane, nitroethane	6.97
2	Mesitylene, *n*-propanol, 2-methyl-2-pentanol, methyl ethyl ketone, 1,4-dioxane, thiophene, nitroethane	7.52
3	Benzene, *n*-butanol, 2-pentanone, nitropropane, pyridine, 2-methyl-2-pentanol, 2-octyne	7.72
4	Ethylbenzene, decalin, ethynylbenzene, 2-propanol, propionaldehyde, 1,2-difluorotetrachloroethane	7.69
5	Ethylbenzene, hydrindane, hexanal, propanol, 1,1-difluoroethane, 2-methyl-2-pentanol	8.06

Recently, several workers have proposed more restricted sets of probes. In Fig. 1, the σ value of the best probe set is represented as a function of the number of probes selected. As might be expected, a continuously descending curve is obtained, so that no distinct preferential number of probes can be deduced from this figure. It also tends to show that there is no reason to choose a certain number of probes *a priori* or to conclude that this number of factors is sufficient to describe the GLC behaviour of a substance. It is rather a question of criteria and how close a description of chromatographic behaviour one wants.

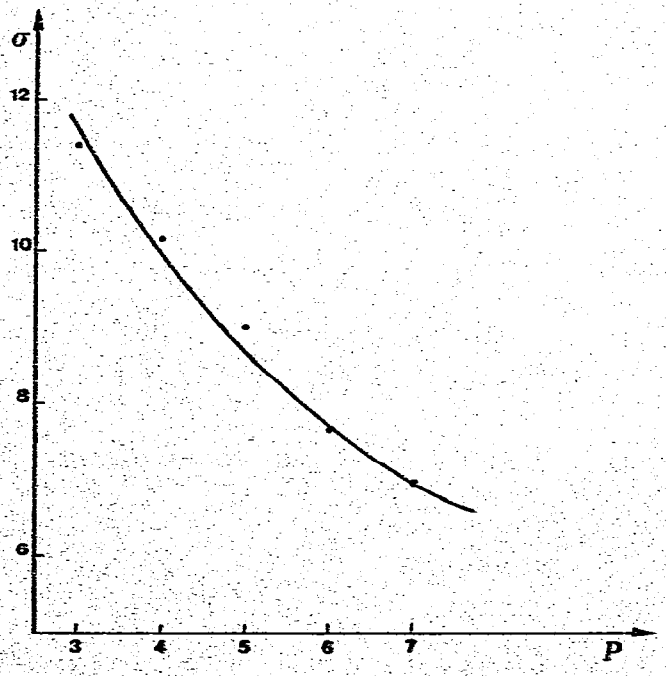

Fig. 1. σ values of the best probe set as a function of the number of probes selected.

Hartkopf and co-workers[10,11] have proposed that only four probes should be used. From considerations of solubility parameters, they concluded: "it points out the redundancy of methyl ethyl ketone as a functional probe". As stated above, there is, according to us, no reason to conclude that a certain number of probes should be used. If one decides that the chromatographic behaviour of the solutes is described well enough with four probes, then one of the five factors represented by the probes must be eliminated. This does not mean, however, that it is redundant, but only that it is less important. As shown in Table II, ethanol, crotonaldehyde, thiophene and dioxane are selected by our procedure from Rohrschneider's set. This means that acetonitrile and propionaldehyde, in the $p = 5$ solution, are replaced with crotonaldehyde. Hartkopf and co-workers[10,11] prefer nitroethane (which closely resembles acetonitrile). The reason for choosing this nitro derivative while the branch and bound procedure selects an aldehyde is that they reasoned that the fourth probe in the set must be a dipole orientator and that it must therefore be a substance whose behaviour is determined as exclusively as possible by this type of interaction, *i.e.* a substance with rather extreme characteristics. The branch and bound procedure selects, from among those substances whose behaviour is determined primarily by dipole orientation forces, the most representative, *i.e.* a substance whose characteristics are less extreme. It is surprising that according to Lowry *et al.*[12], the probe which should be eliminated first from the set of five is benzene. This is in disagreement with both our results and those obtained by Hartkopf and co-workers[10,11]. It is perhaps due to the criterion chosen by Lowry *et al.*, based on the ability of the probes to characterize liquid phases according to their distance from a number of standard phases.

In Tables VI and VII, the predicted results for several sets of four probes are given. Sets 3 and 10 in Table VI were taken from Ta[illegible]IV in the paper by Hartkopf *et al.*[11], set 3 being his best probe set. Sets 7 and 9 were taken from the paper by Lowry *et al.*[12]. Note that set 9 is Lowry *et al.*'s best choice but that, in fact, the other set (with benzene) is found to be better (although only slightly so). Sets 5, 6 and 8 are sets with the more usual probes, tried by us in an effort to develop a probe set which would be a better alternative than the probe sets proposed by the other workers. Set 1 is the set selected from McReynolds' set by our procedure (see also Table I) and set 2 is that obtained from Table II. Set 4 was obtained from McReynolds' set

TABLE VI

PREDICTION WITH FOUR PROBES IN McREYNOLDS' SET

Rank number	*Probes*	σ
1	2-Propanol, propionaldehyde, 1,1-difluorotetrachloroethane, thiophene	10.35
2	Ethanol, crotonaldehyde, thiophene, dioxane	10.50
3	Nitroethane, *n*-propanol, benzene, dioxane	10.83
4	Ethylbenzene, ethynylbenzene, 2-propanol, propionaldehyde	10.91
5	*n*-Butanol, 2-pentanone, benzene, pyridine	11.03
6	Ethanol, benzene, methyl ethyl ketone, pyridine	11.61
7	Benzene, *n*-butanol, nitropropane, pyridine	11.72
8	Benzene, pyridine, acetone, ethanol	11.80
9	*n*-Butanol, 2-pentanone, nitropropane, pyridine	11.92
10	Nitroethane, chloroform, benzene, dioxane	12.00

TABLE VII

PREDICTION WITH FOUR PROBES IN ROHRSCHNEIDER'S SET

Rank number	*Probes*	σ
1	Ethanol, crotonaldehyde, thiophene, dioxane	7.59
2	Nitroethane, *n*-propanol, benzene, dioxane	7.71
3	Benzene, pyridine, acetone, ethanol	8.50
4	Ethanol, benzene, methyl ethyl ketone, pyridine	8.62
5	2-Propanol, propionaldehyde, 1,1-difluorotetrachloroethane, thiophene	9.60

using the $p = 6$ solution in Table I, after elimination of two probes representative of less than three substances.

There can be no doubt in this instance that the results obtained by the branch and bound method are consistently better than those obtained by chemical reasoning. This statement should not be taken as an invitation to stop reasoning in this way, but only as an indication that a "chemically blind" mathematical method can serve as a guide in such theoretical work.

It would be tedious to examine Table VII in the same detailed way and it will suffice to note that the best result is that obtained using the $p = 6$ result from Table II after elimination of two probes for the reasons explained above.

If only three probes are used, then according to Table II they should be benzene, crotonaldehyde and cyclopentanol. Benzene is now representative of the typical dispersion force probes and pyridine, dioxane and other probes which have been identified as proton acceptors. This set is analogous to the set of three chosen from McReynolds' data set, *i.e.* 2-propanol, crotonaldehyde and thiophene. Our conclusion is in agreement with that of Lowry *et al.* in the sense that from their Table I one observes that the best results are obtained with an alcohol, a ketone or a nitro derivative and an aromatic compound or pyridine. However, only the set benzene, *n*-butanol and nitropropane or 2-propanone yields a prediction result comparable to the probe set proposed by us. The set benzene, *n*-butanol, 2-pentanone even provides the best result ($\sigma = 11.42$). The sets containing pyridine instead of benzene yield much poorer results (sets 9–11). The sets 2–7 were obtained by the branch and bound method using either Rohrschneider's or McReynolds' set, $(1-\varrho)$ or $\sqrt{1-\varrho}$ as the similarity parameter and $p = 3$, 4 or 5 results (the latter two after elimination of one or two probes). Therefore, one observes again that the prediction is usually better with branch and bound methods than with literature results (Lowry *et al.* proposed as the best probe choices sets 10, 9 and 8 of Table VIII in that order, set 1 being only their fourth choice.)

By comparing the classification of the substances in Rohrschneider's set according to the nearest probe for $p = 5$, 4 and 3 in Table III, one observes that the classification for $p = 4$ is simply obtained by grouping the propionaldehyde and acetonitrile classes in the $p = 5$ classification. When going from $p = 4$ to $p = 3$, all of the solutes in the dioxan class come together with most of the thiophene-class solutes. Two compounds from the latter are relocated, however: phenylacetylene and difluorotetrachloroethane are now found together with the alcohols. This seems to indicate that classification with $p = 3$ becomes uncertain and that $p = 4$ is therefore the number of probes where the compromise between experimental convenience

TABLE VII

PREDICTION RESULTS WITH THREE PROBES IN McREYNOLDS' SET

Rank number	*Probes*	σ
1	Benzene, *n*-butanol, 2-pentanone	11.42
2	2-Propanol, propionaldehyde, thiophene	11.73
3	Ethylbenzene, 2-propanol, hexanal	12.01
4	2-Propanol, crotonaldehyde, thiophene	12.03
5	Ethanol, crotonaldehyde, thiophene	12.04
6	Styrene, 3-hexanol, propionaldehyde	12.35
7	Benzene, crotonaldehyde, cyclopentanol	12.61
8	Benzene, *n*-butanol, nitropropane	13.56
9	*n*-Butanol, 2-pentanone, pyridine	14.10
10	*n*-Butanol, nitropropane, pyridine	15.81
11	*n*-Butanol, pentanone, nitropropane	16.63

and closeness of fit of the description is optimal. The results indicate that the sets proposed by us are often better than those already proposed in the literature. We should state explicitly that this does not mean that we propose that the use of existing probe sets should be discontinued; the existing probe sets are rarely much worse and, in any case, the same factors (proton donors, dispersion force, etc.) are selected. The choice of the actual probes, moreover, is clearly dependent on the data set used. Also, we think that it would be preferable if a group of experts were to decide finally what probes should be used for the general characterization of stationary phases.

As stated by Haken[20], there is little value in the continued introduction of general schemes which are essentially similar; we have included the word "general" here because we think that it could be very interesting to develop probe schemes for more restricted groups of compounds such as *e.g.* lipids. This would than allow classification and selection of preferred phases for analysis of this restricted group of compounds. A combined information theoretical–numerical taxonomy selection procedure proposed by us in collaboration with Eskes *et al.*[21] leads to different preferred phase sets for groups such as aldehydes and ketones or alcohols. We concluded, therefore, that there is little point in trying to develop a small set of preferred phases for all GLC uses. It cannot be denied that there is a large redundancy in GLC phases but, at the same time, one cannot reasonably hope to achieve all GLC separations with a restricted general set of phases. This conclusion is also in agreement with the findings of Haken[13], who noted that general preferred phase sets are of little value in a field such as lipid analysis. We would therefore like to complement our conclusion in the earlier paper[21] by proposing to reduce the number of existing phases by examining the major specialized areas of application of GLC (such as lipid analysis) and developing preferred phase sets for those areas.

In the same way as general preferred phases are often of no value in specialized fields, it also seems preferable to select special probes for those fields. Until now, the selection of probe sets for specialized fields of GLC has not been undertaken, because of the difficulty of selecting such probes by chemical reasoning. It should be noted that the selection procedure used here does not presuppose any knowledge of the physico-chemical nature of the substances so that it should be of value for the selection of functional probes in such specialized areas of GLC. Those who are deterred by

the rather involved mathematics of the branch and bound methods should be reminded that the results obtained by the very simple heuristic method are nearly as good*.

This plea for specialized probes should not be mistaken as an invitation to abandon general probes. It is obvious that these are necessary to unify the field of GLC and that both a general characterization and a general classification of phases is an important first step in the selection of these phases. Moreover, the probe concept has been, and still is, of great value in understanding the interactions between solutes and stationary phases and it is in this respect that the work of the authors cited is important and worthwhile.

Finally, it should be noted that the application of these methods is, of course, not limited to GLC and that it should be useful in other areas of separation science. More generally, operations research provides a number of hitherto insufficiently used methods of great potential value in analytical chemistry.

ACKNOWLEDGEMENTS

The authors thank W. O. McReynolds for placing his data at their disposal and F.K.F.O. for financial assistance.

REFERENCES

1 D. L. Massart and L. Kaufman, *Anal. Chem.*, 47 (1975) 1244A.
2 L. Rohrschneider, *J. Chromatogr.*, 22 (1966) 6.
3 W. O. McReynolds, *J. Chromatogr. Sci.*, 8 (1970) 685.
4 J. J. Leary, J. B. Justice, S. B. Tsuge, S. R. Lowry and T. L. Isenhour, *J. Chromatogr. Sci.*, 11 (1973) 201.
5 S. Wold and K. J. Andersson, *J. Chromatogr.*, 80 (1973) 43.
6 D. L. Massart, P. Lenders and M. Lauwereys, *J. Chromatogr. Sci.*, 12 (1974) 617.
7 L. Rohrschneider, *Anal. Chem.*, 45 (1973) 1241.
8 L. R. Snyder, *J. Chromatogr.*, 92 (1974) 223.
9 D. L. Massart and H. de Clercq, *Anal. Chem.*, 46 (1974) 1988.
10 A. Hartkopf, *J. Chromatogr. Sci.*, 12 (1974) 113.
11 A. Hartkopf, S. Grunfeld and R. Delumeya, *J. Chromatogr. Sci.*, 12 (1974) 119.
12 S. R. Lowry, S. Tsuge, J. J. Leary and T. L. Isenhour, *J. Chromatogr. Sci.*, 12 (1974) 124.
13 J. K. Haken, *J. Chromatogr. Sci.*, 13 (1975) 430.
14 C. van de Panne, *Linear Programming and Related Techniques*, North Holland, Amsterdam, London, 1971.
15 H. A. Land and A. G. Doig, *Econometrica*, 28 (1954) 497.
16 L. Kaufman, *Ph. D. Thesis*, University of Brussels, 1975.
17 W. O. McReynolds, unpublished results.
18 J. J. Leary, S. Tsuge and T. L. Isenhour, *J. Chromatogr.*, 82 (1973) 366.
19 *Applied Science Catalog.*, Applied Science Labs., State College, Pa., U.S.A., No. 18, 1975.
20 J. K. Haken, *J. Chromatogr.*, 73 (1972) 419.
21 A. Eskes, F. Dupuis, A. Dijkstra, H. de Clercq and D. L. Massart, *Anal. Chem.*, 47 (1975) 2168.

* Fortran IV programs can be obtained from the authors.