Analytica Chimica Acta, 157 (1984) 227—243
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

AN AUTOMATED LIBRARY SEARCH SYSTEM FOR ^{13}C-N.M.R. SPECTRA BASED ON THE REPRODUCIBILITY OF CHEMICAL SHIFTS

R. W. BALLY, D. VAN KRIMPEN, P. CLEIJ and H. A. VAN 'T KLOOSTER*

State University of Utrecht, Laboratory for Analytical Chemistry, Croesestraat 77A, 3522 AD Utrecht (The Netherlands)

(Received 2nd August 1983)

SUMMARY

A library search system for ^{13}C-n.m.r. spectra, based on a statistical description of the reproducibility of chemical shifts, is presented. A similarity index in the form of a significance probability (*P*-value) is developed from a previously introduced general concept. The applied data base of some 6000 spectra originates from the NIH-EPA Chemical Information System (CIS). The reproducibility model and the retrieval system are developed on a CDC Cyber 175 computer, with PASCAL as the programming language. The performance of the system is evaluated by using recall/reliability and confusion/recall plots. In a tentative comparison with the Clerc search method (included in the CIS package), the Utrecht ^{13}C-n.m.r. reproducibility-based retrieval system shows a better identification performance. The system is adaptable for use on a microcomputer.

In recent years, several systems have been proposed for computer-aided library search of ^{13}C-n.m.r. spectra, aiming at the identification of organic compounds and based on various coding and comparison algorithms [1—13]. Collections of several thousands of spectra, from different sources (i.e., recorded on different instruments, under various experimental conditions) are usually applied as data bases. Such data bases, which are often commercially available, generally suffer from poor interlaboratory reproducibility of the spectral data involved. In a previous paper [14], a new similarity index for "straightforward" library search methods was introduced, primarily for application to this type of data base. The main object of straightforward search methods is to retrieve the reference data (if available) of the unknown compound. This contrasts with "interpretative" methods, which aim principally at retrieving data of compounds similar in structure to the unknown. The proposed index has the form of a significance probability and is developed from a statistical model of the reproducibility of the quantities used for the comparison of unknown and reference data. Defined in general terms, this index is applicable to different types of analytical data, provided that the variables used to characterize unknown and reference spectra (feature quantities) are continuous in nature. In ^{13}C-n.m.r. spectra, the chemical shift and the peak intensity are such variables, in contrast to multiplicities. When a large multisource data base such as the CIS collection is used, how-

0003-2670/84/$03.00

ever, peak intensities are specified only in a minority of the spectra and show very poor reproducibility, which makes this information unsuitable for library search.

This paper reports the development and evaluation of a straightforward ^{13}C-n.m.r. library search system based on chemical shifts as the features. As a criterion of matching, a similarity index is developed based on a statistical description of the reproducibility of chemical shifts. The system thus takes into account any systematic and random errors occurring in chemical shifts. As a first step in the development of the retrieval system, a reproducibility model was derived from some 200 pairs of "alternative spectra"; each pair was for the same compound, but the spectra were recorded under different experimental conditions. This model describes the statistical behaviour of differences in chemical shifts in these alternative spectra. The resulting reproducibility function forms the basis of the definition of the similarity index (S_I). The main function of this index is to permit a separation of the reference compounds into two classes, i.e., compounds that could be the unknown and compounds that cannot. This corresponds to the acceptance or rejection of the null hypothesis that the unknown and the reference compound are identical. By specifying a threshold value of the index, all references with an index value above this threshold should be retrieved as belonging to the "possible" class. The precise value of this threshold depends on the pre-accepted risk of misclassifying (not retrieving) the correct reference compound. The result is a list of reference compounds (if any), one of which could be identical to the unknown, instead of the usual 5—10 "best" matches (which may in fact be very bad matches). The main features in the development of the ^{13}C-n.m.r. reproducibility-based retrieval (C13RR) system are schematically indicated in Fig. 1.

EXPERIMENTAL

For the design, development and evaluation of the C13RR system, a CDC Cyber 175 (60 bits) computer was used. Programs are written in PASCAL, with the exception of a few subroutines; a Wiswesser line notation/connectivity table conversion program (kindly supplied by Prof. J. Zupan, Kemijski Institute Boris Kidrič, Ljubljana, Yugoslavia) was written in FORTRAN. The data base originated from the NIH-EPA Chemical Informa-

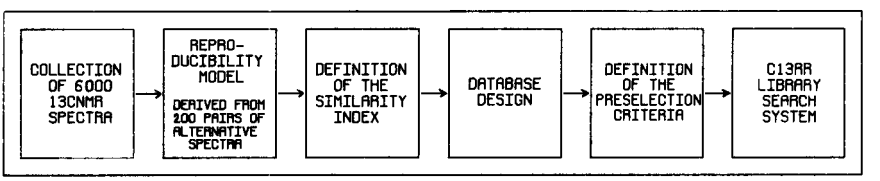


Fig. 1. Main features in the development of the ^{13}C-n.m.r. reproducibility-based retrieval (C13RR) system.

tion System (CIS; 1982 version) and contained some 6200 spectra of 5800 different compounds. The original data base, coded in Bremser exchange format [7], was transformed and reorganised, resulting in B-tree-structured and indexed sequential subfiles [15]. An important adaptation was the removal of redundant shift values; the original data base contained multiple shift values of equal magnitude, obviously referring to the same peak. In the data base ultimately applied, a chemical shift was stored with a precision of 0.01 ppm.

A REPRODUCIBILITY-BASED SIMILARITY INDEX

The similarity index proposed by Cleij et al. [14] implies the use of so-called difference quantities, representing the differences in value, for two spectra, of feature quantities. The use of chemical shifts as the feature quantities leads to a set of n difference quantities, $\Delta q_1 \ldots \Delta q_n$, defined by

$$\Delta q_i = \delta_{U,i} - \delta_{R,i} \quad (\text{for } i = 1 \ldots n) \tag{1}$$

where $\delta_{U,i}$ and $\delta_{R,i}$ are the shift values of the ith peak in the unknown and reference spectrum, and n is the number of peaks (assuming an equal number of peaks in both spectra).

Reproducibility of chemical shifts

For the development of the model of reproducibility of chemical shifts, a set of 202 pairs of alternative spectra (each pair for the same compound but consisting of different spectra), originating from the CIS data base, was used. This set was selected from 441 such pairs, by exclusion of pairs of spectra which were identical, or contained less than 3 peaks, or had different numbers of peaks. The 202 pairs of alternative spectra were analysed in the form of "difference spectra". A difference spectrum of two spectra A and B is a plot of the differences in shift values for corresponding peaks, versus the shift values of spectrum A, as shown in Fig. 2.

The differences in shift values, as shown by the difference spectra, consist of a random and a systematic part. The random part is demonstrated by the scatter of the data points in a difference spectrum. The magnitude of this scatter ranges from very small (Fig. 3), to relatively large (Fig. 4). The systematic part of the differences in shift values is demonstrated by difference spectra in which the average difference significantly deviates from the zero (no difference) line (Fig. 5). Evaluation of some 50 difference spectra showed that systematic differences can adequately be described by a straight line, parallel to the zero line. These systematic deviations of the shifts are mainly caused by variation in the (absolute) position of the TMS peak.

The reproducibility function, for the set of n difference quantities denoted by $p_0[\Delta q_1 \ldots \Delta q_n]$, should provide a description of the formal difference spectra. As shown above, the total difference in a chemical shift (Δq_i) is considered to consist of the random part (Δq_i^r) and the systematic part (Δq^s), i.e.:

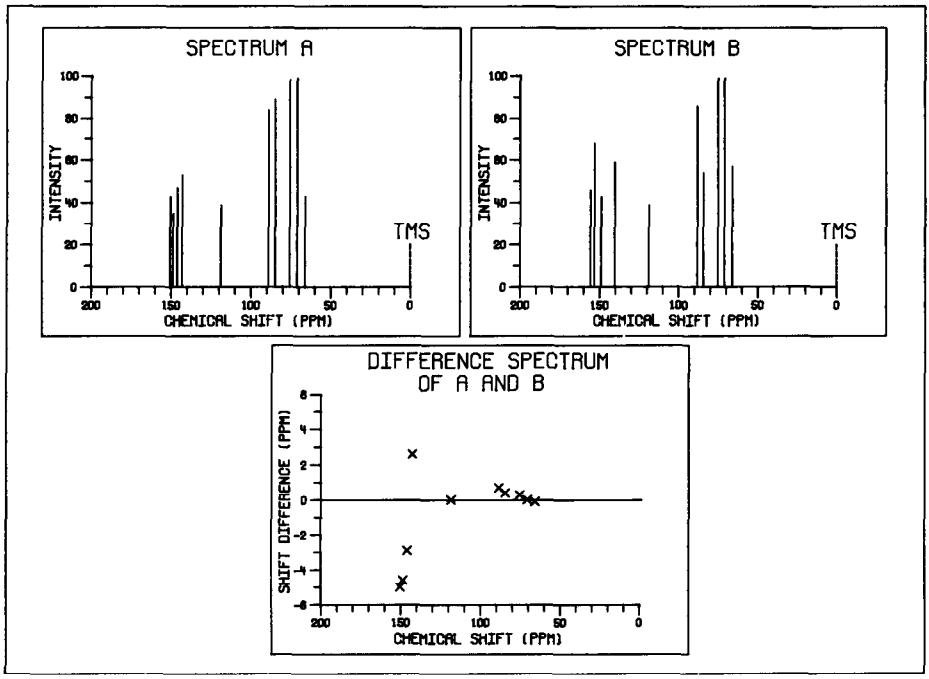


Fig. 2. Two ^{13}C-n.m.r. spectra A and B of the disodium salt of adenosine *s*'-(tetrahydrogentriphosphate), dissolved in two different basic solutions [16]. The "difference spectrum" of A and B comprises the differences in the shift values of the peaks in spectrum A and B.

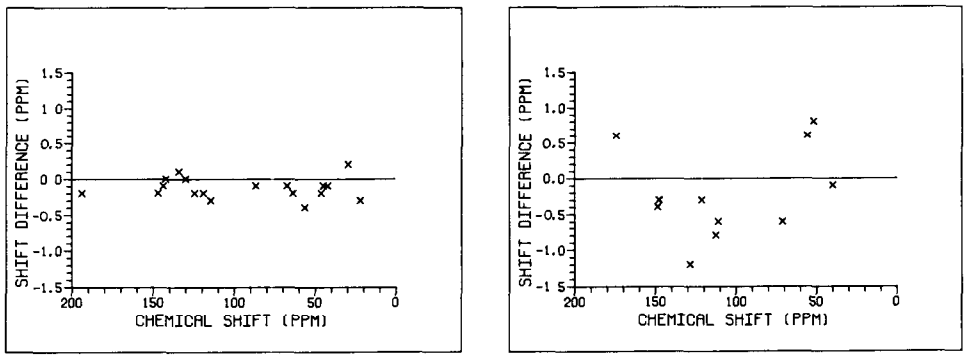


Fig. 3. Example of a difference spectrum, illustrating random differences with a relatively small variance. Differences in the shift values of two spectra of 3,14-bis(acetyloxy)-7,8-didehydro-4,5-epoxy-17-methyl-, (5-α)-morphinan-6-one. The two spectra were measured on different spectrometers [17, 18].

Fig. 4. Example of a difference spectrum, illustrating random differences with a relatively large variance. Differences in shift values of two spectra of benzenepropanoic acid, α-hydroxy-3,4-dimethoxymethyl ester. The two spectra were measured on different spectrometers and in different solvents [19].

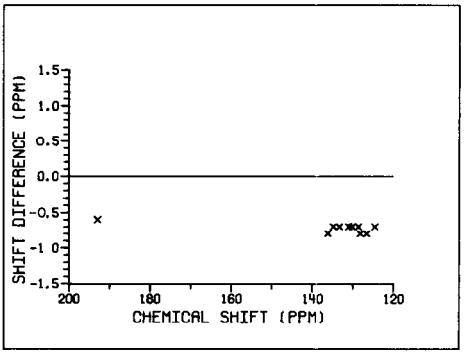


Fig. 5. Example of a difference spectrum, demonstrating a systematic difference in the shift values of two spectra of 1-naphthalenecarboxaldehyde, measured by two different operators [16, 20].

$$\Delta q_i = \Delta q_i^r + \Delta q^s \tag{2}$$

This is illustrated by Fig. 6.

The random differences are assumed to be normally distributed about zero and with variance σ^2. Hence

$$p(\Delta q_i^r) = N(\Delta q_i^r{:}0, \sigma^2) \tag{3}$$

[$N(x:E, V)$ is used in this paper as the general notation for a normal density function with variable x, expected value E and variance V.]

In order to meet the large variations in the magnitude of the scatter observed in difference spectra, the variance σ^2 is considered as a stochastic quantity, varying with the difference spectrum. Preliminary studies showed that the probability density of σ^2 is adequately described by a log-normal function:

$$p(\sigma^2) = [\sigma^2 \cdot (2\pi \cdot V \ln \sigma^2)^{1/2}]^{-1} \exp[-0.5(\ln \sigma^2 - E \ln \sigma^2)^2 / V \ln \sigma^2] \tag{4}$$

where $E \ln \sigma^2$ is the expected value of $\ln \sigma^2$ and $V \ln \sigma^2$ is the variance of $\ln \sigma^2$.

The systematic difference is a constant in a particular difference spectrum but a variable for various difference spectra. Assuming that the systematic difference follows a normal probability function with the same variance as applied to the random differences, the probability distribution of the systematic differences can be described by:

$$p(\Delta q^s) = N(\Delta q^s{:}0, \sigma^2) \tag{5}$$

For a set of n difference quantities (differences in shift values) the reproducibility function can then be formulated as

$$p_0[\Delta q_i] = \int_{\sigma^2=0}^{\infty} p(\sigma^2) \cdot \int_{\Delta q^s=-\infty}^{\infty} p(\Delta q^s) \cdot \prod_{i=1}^{n} p(\Delta q_i^r)\, d\Delta q^s d\sigma^2 \tag{6}$$

in which n is the number of peaks of the spectra involved, $[\Delta q_i]$ is the set of n difference quantities, $p(\sigma^2)$, $p(\Delta q^s)$, $p(\Delta q_i^r)$ are given by Eqns. 4, 5 and 3, and where $\Delta q_i^r = \Delta q_i - \Delta q^s$ (see Eqn. 2). An elaboration of Eqn. 6 is given in the Appendix.

The similarity index (S_I)

The similarity index [14] is defined as:

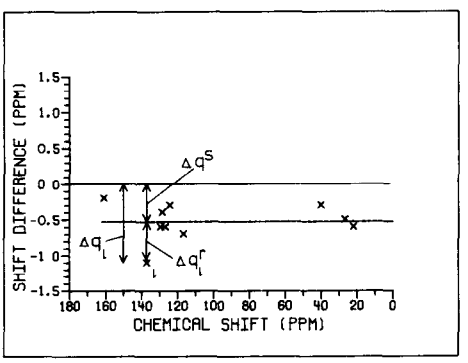
$$S_I = \int_{\Delta q_1} \dots \int_{R[\Delta Q_i]} \int_{\Delta q_n} p_0[\Delta q_i]\, d\Delta q_1 \dots d\Delta q_n \qquad (7)$$

where n is the number of chemical shifts, $p_0[\Delta q_i]$ is given by Eqn. 6 and $R[\Delta Q_i]$ is the region in the space of difference quantities, defined by the condition $p_0[\Delta q_i] \leqslant p_0[\Delta Q_i]$, with ΔQ_i being the "observed" value of Δq_i, i.e., the difference of the shift values of the ith peak in the unknown and reference spectrum.

An elaboration of Eqn. 7 into a more workable expression of the S_I is given in the Appendix.

THE RETRIEVAL SYSTEM

The ^{13}C-n.m.r. reproducibility-based retrieval (C13RR) system consists of a data base, the (modular) set of search programs and job control software of the computer being used. In the standard mode, the system retrieves



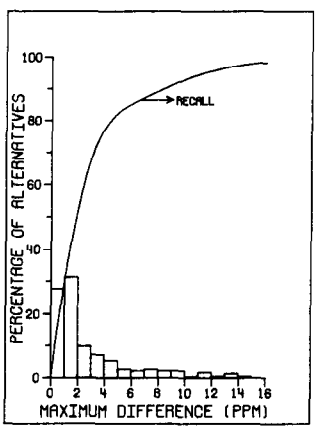


Fig. 6. The total difference (Δq_i) is the sum of the random difference (Δq_i^r) and the systematic difference (Δq^s). This difference spectrum results from two spectra of 3,4-dihydro-1(2H)-quinolinecarboxaldehyde, measured in different solvents [21].

Fig. 7. Histogram of the maximum differences in shift values as occurring in 230 pairs of alternative spectra. The curve represents the theoretical recall as a function of the maximum difference. For 15 ppm (default value), the recall is 98%.

all the reference spectra with a number of peaks equal to that of the unknown spectrum and with an S_I value higher than the threshold level, usually preset at 2%. In order to deal with "missing" peaks in ^{13}C-n.m.r. spectra, an alternative mode, allowing a difference of one peak between unknown and reference spectrum, is optional. (Missing peaks are mainly caused by differences in instrumental and recording conditions, sample impurities and careless copying of spectral data.) Table 1 gives a survey of missing peaks in 316 pairs of alternative spectra, from which it can be concluded that more than 1 peak is missing in only 9% of the spectra. The fact of a missing or extra peak in the reference spectrum is neglected in the calculation of the S_I. In the alternative mode, three kinds of hit lists are produced for an unknown spectrum with n peaks, i.e., lists containing reference spectra (if any) with $n-1$, n and $n+1$ peaks.

Preselections

If the algorithm for the calculation of the S_I were applied to all the reference spectra, this would lead to an inacceptably large response time, especially with large data bases. A way to solve this problem is the use of preselection criteria, which should meet the following requirements: (1) all reference spectra with an S_I greater than the applied threshold level should pass the preselection; (2) the preselection should have a high total selectivity; (3) the (computer) time consumption should substantially be decreased.

In the C13RR system the following preselection criteria are applied sequentially. First, in the standard mode, the number of peaks in the unknown and the reference spectra must be equal; in the alternative mode a difference of one peak is allowed. Secondly, the maximum allowable difference in corresponding shift values is 15 ppm. The second criterion was derived from an evaluation of the set of pairs of alternative spectra. The distribution of the maximum differences between corresponding shift values occurring in the pairs of alternative spectra is given in Fig. 7; it may be interpreted as the recall [22] as a function of the tolerance. A tolerance of 15 ppm (the default value in the system) corresponds to 98% recall.

TABLE 1

Frequency of differences in the total number of peaks for 316 pairs of alternative spectra

Difference in number of peaks	Pairs of alternative spectra	Percentage of total
0	230	72.8
1	57	18.0
2	15	4.8
3	8	2.5
4	4	1.3
5	1	0.3
6	1	0.3
>6	0	0

Organisation of the data base

The data base is organised according to the preselection criteria. An outline of the data base is given in Fig. 8. As shown, the original data base is divided into two parts: the first contains the spectral data, and the second holds all further information such as names, molecular formulae, CAS registration numbers, sources of origin, etc. This division is made because the retrieval algorithm only uses the shifts of the spectra. The part containing the shifts is divided into subfiles, each containing spectra with the same number of shifts. Within a subfile the spectra are ranked according to the highest (most significant) shift values of the spectra. The reason for this ranking is the use of an indexed sequential file organisation [15], which we consider to be an efficient access strategy in combination with the second preselection criterion. The other part of the data base consists of index files on CAS registration number and molecular formula, which allows the

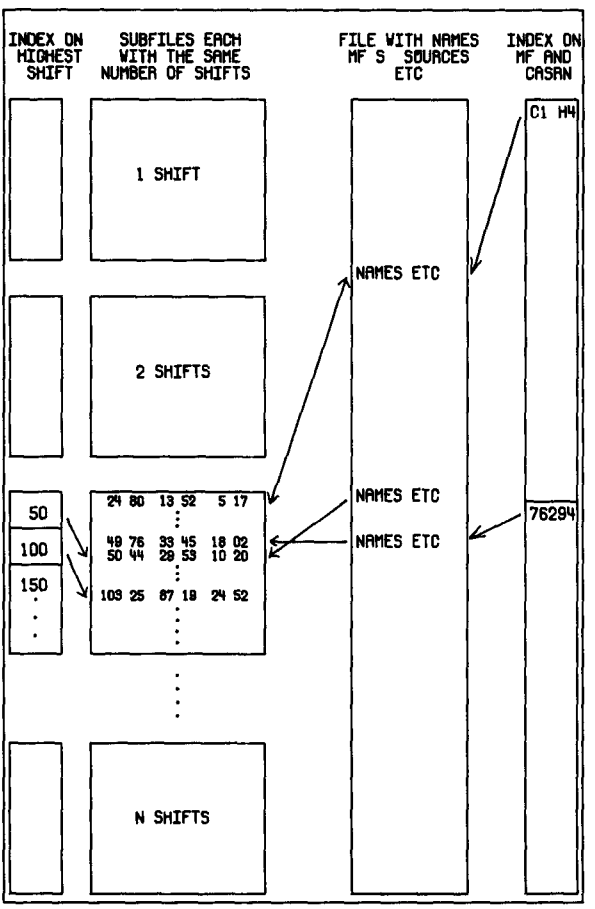


Fig. 8. Organisation of the data base used in the C13RR system.

establishment of the presence or absence of a specified compound as a reference in the data base. These files contain a "dense" index and are organised as a B-tree [15] with 128 keys per node, resulting in at most two disk accesses in order to find all the keys that belong to a given entry.

The retrieval algorithm

The main steps of the retrieval algorithm are as follows.

1. The subfile corresponding with the number of shifts in the unknown spectrum is selected.
2. On this subfile a window is defined by the highest shift in the unknown spectrum ± the preset tolerance.
3. The selection starts with the first spectrum that is "seen" through the window. The position of this spectrum in the subfile is calculated with the use of the index file. From this point the sequential access plan is executed.
4. Each difference between a shift of the unknown and the corresponding shift of the reference spectrum is compared with the applied tolerance; a spectrum is discarded when a difference greater than the tolerance is found.
5. If a reference spectrum has not been discarded in step 4, the S_I is calculated. The reference spectrum is selected for the hit list if the S_I value is greater than the applied threshold value. A binary tree is used to rank the reference spectra according to the S_I values.
6. If the highest shift in a subsequent reference spectrum falls within the specified window, it is submitted to step 4, etc., otherwise the program terminates and the results of the search are displayed.

As mentioned above, it is possible to allow a difference of one peak in the total number of peaks of the unknown and the reference spectrum. In this case, the retrieval algorithm deletes from the "larger" spectrum a shift in such a way that the maximum S_I value for that reference is calculated.

Modular structure of the C13RR system

The C13RR system consists of five modules. The program starts in the module CONTROL, which contains the system commands and controls the input and storage of the unknown spectra. Options for further processing of the input data are activated by specifying the first character or the whole option label. The user-friendly operation of the system includes the display of warnings after wrong statements have been typed. Typical outputs of each module are illustrated in Fig. 9.

EVALUATION OF THE C13RR SYSTEM

An essential criterion in the evaluation of a retrieval system is the discriminating power of the comparison index in separating the target spectrum (spectra) from the remaining reference spectra. Two aspects of this property can be distinguished, viz., the "absolute" and the "relative" discriminating power. A good absolute behaviour is demonstrated when the distribution of

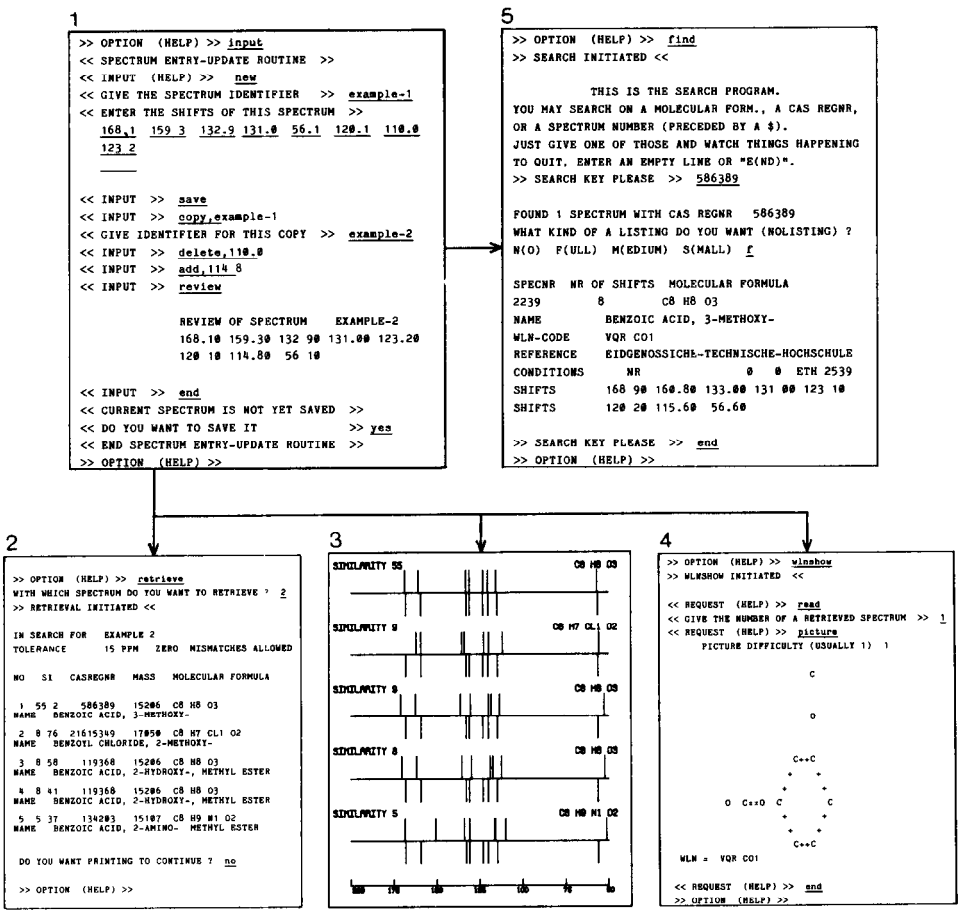


Fig. 9. Typical outputs of the C13RR system modules. The system starts in the module CONTROL, which allows the operator to enter the spectral information of the unknown. Included are several options such as: add and delete a shift, copy and review a whole spectrum or save a spectrum. In CONTROL it is also possible to select one of the other modules. RETRIEVE (2): the system starts the library search and presents the result of the search in the form of a hit list, containing reference spectra (if available) with a similarity greater than the specified threshold (default 2%); a typical response time for one search is 10 s. PLOT (3): the system produces plots of the unknown spectrum (pointed downwards) versus the retrieved reference spectra (pointed upwards), allowing a visual comparison. WLN-SHOW (4): the system produces a structure formula from a Wiswesser line notation [23] of a compound present in the hit list. FIND (5): after entering a search key (CAS reg. no. or molecular formula), the system gives a list of reference compounds containing the search key.

index values for correct matches (reference compound = unknown compound), determined for a large set of test spectra, shows only a small overlap with the distribution for incorrect matches. In the case of an optimal absolute behaviour of the comparison index, it is possible to define a single threshold value that divides all matches (for all searches) into correct and

incorrect ones (or correct positives and false positives), respectively. A retrieval system with a good relative discriminating power generally assigns, within one search, a "better" index value to the target spectrum (or spectra) than to the remaining reference spectra. An optimal system in this respect always places the target spectrum (or spectra) on top of the list of retrieval results. Such a system, however, is not necessarily an optimal or even a good retrieval system with respect to the absolute behaviour of the comparison index. In contrast, a comparison index with an optimal behaviour in the absolute sense, also has an optimal relative behaviour. It may seem, at first sight, that a good relative discriminating power of the comparison index is a sufficient guarantee for an adequate performance of the retrieval system. This is, however, not wholly true, especially not with regard to an important property of a good retrieval system, i.e., the possibility for the user of the system to judge, on the basis of the index values, whether the reference listed first is correct, or that no reference spectrum is available for the unknown compound. When a comparison index with a good discriminating power is used in the absolute sense, it is relatively easy to distinguish between the case of the first retrieved reference being the target spectrum, and the situation in which no target spectrum is available. With a comparison index that demonstrates a good relative behaviour, yet shows poor performance as an absolute index, these situations are much more difficult to distinguish.

For these reasons the absolute discriminating power of the similarity index is chosen as the main criterion in the evaluation of the C13RR system. The relative behaviour of the S_I, however, is also considered, because the relative and absolute behaviours are correlated only to a certain extent.

The absolute discriminating power of a comparison index can be evaluated with recall/reliability plots [22]. This type of evaluation fits with a retrieval plan in which all reference spectra having an index value above (or below) some prespecified threshold value, are retrieved. The recall for a test set of "unknown" spectra is defined as the number of actually retrieved target reference spectra (correct positives), divided by the total number of target spectra, available for the test set. Provided that the unknown compound has a spectrum in the reference file, the recall represents the estimated probability that this target spectrum is actually retrieved. The relationship between the recall and the threshold value of the S_I [14] is Recall = $(1 - Th)$.

The reliability is defined as the number of retrieved target spectra, divided by the total number of retrieved reference spectra (correct and false positives) for the "unknown" spectra of the test set. Both the recall and the reliability are functions of the index threshold value. Increasing the threshold value of the similarity index will decrease the recall and increase the reliability. For a good relative behaviour, the increase in reliability will exceed the decrease in recall. A recall/reliability plot can thus be obtained by varying the threshold values and plotting the reliability values as a function of the recall values. Recall/reliability values can, in principle, be calculated from the distributions

of the index values for correct and false positives, which explains their utility in evaluating the absolute behaviour of comparison indices.

A variation of the recall/reliability evaluation method is the extension of the concept "correct" for the evaluation of the reliability, by considering a retrieved reference spectrum not only as correct if the reference compound is identical to the test compound but also if this reference compound is, within well-defined limits, similar in structure to the test compound. This method of evaluation has been applied to the PBM retrieval system for mass spectra, using four definitions of a similar compound [2]. The performance of a retrieval system with regard to the relative behaviour of the comparison index, should be evaluated on the basis of the ranking of correct and incorrect matches for individual searches, executed with the spectra of the test set. As a measure of this behaviour, the recall as a function of the "confusion" [24] has been used. The confusion is defined here as the number of false positives with a S_I equal or better than the S_I of the target reference spectrum.

The C13RR system was compared with the CIS/Clerc CNMR search system [2, 7] with regard to the relative behaviour of the comparison index. (A recall/reliability plot of the Clerc system is not feasible, because of the nature of the applied matching criterion.) The (commercially accessible) CIS/Clerc system was chosen, as it has the same data base and is also meant to be used as an aid in the identification of organic compounds.

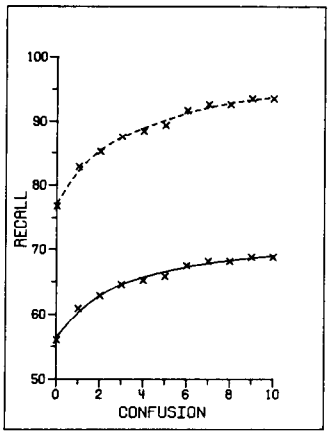
Test spectra

A test set (A) of 264 spectra was selected from the data base, each having one or more alternative spectra (no exact duplicates) as target spectra in the data base. For the whole set, 304 target spectra were available. All 264 test spectra were processed using the complete reference file of 6200 spectra minus the test spectra. To establish the influence of "missing" peaks, a second test set (B) was composed of 194 spectra having only target spectra in the reference file with the same number of peaks.

For a comparison with the CIS/Clerc system a third test set (C) of 50 spectra was randomly selected from test set A.

Results

The confusion plots of test sets A and B, given in Fig. 10, clearly show the negative effect of missing peaks. In processing the spectra of test set A, both preselection criteria are effective, resulting in a maximum attainable recall of 71% (27% of the correct positives are "lost" in the first, 2% of the remaining in the second preselection). The plot for test set B has a maximum of 94%, which is close to the limit of 98%, caused by the second preselection criterion (the first one has no effect in this case). From Fig. 10 it can be concluded that the correct positive is generally among the first retrieved references, or it is not retrieved at all. The recall/reliability plots of test sets A and B, for threshold values of the S_I ranging from 90% to 2%, are given in Fig. 11. Both curves can be divided into two regions of the recall, i.e., a first



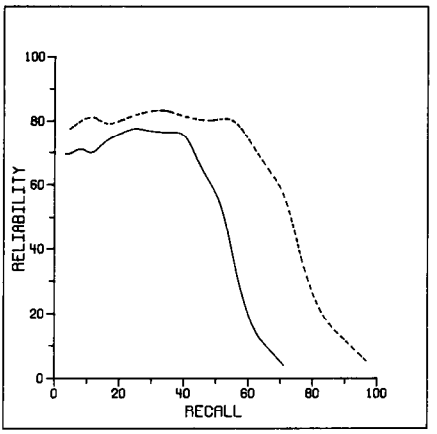
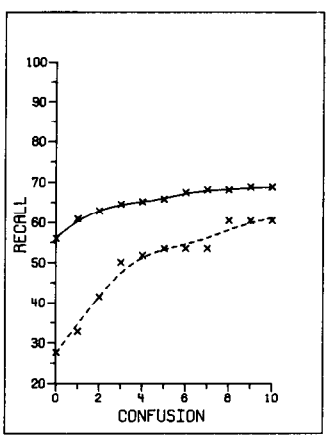


Fig. 10. Confusion plots for two test sets. Test set B (– – –) consisting of 194 spectra with target spectra having equal numbers of peaks. Test set A (——) containing test set B plus 70 spectra with target spectra having different numbers of peaks. (cf. Table 1.)

Fig. 11. Recall/reliability plots of test set A (——) and test set B (– – –). (cf. Fig. 10.)

region with about constant reliability (the plateau) and a region with strongly decreasing reliability. The plateau can be explained by the fact that there are some sets of compounds with (in principle) identical spectra. For instance, if it is assumed that (a) a reference file of 10 000 spectra consists of 5000 pairs of compounds, (b) the spectra of each pair differ only within measur-



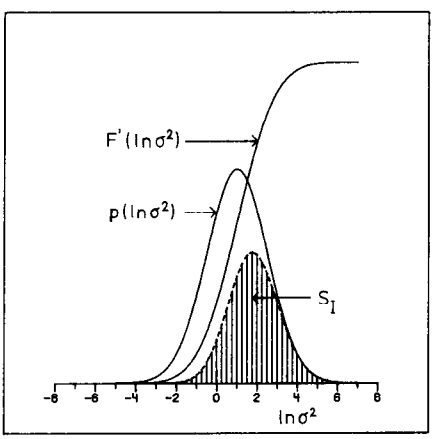


Fig. 12. Confusion plots of the C13RR system (——) and the CIS/Clerc CNMR search system (– – –).

Fig. 13. Graphic representation of the similarity index and related functions, according to Eqns. A14–16.

ing errors, and (c) spectra of different pairs are significantly different, this would result in a reliability of 50% for the full recall range. The tail of the recall/reliability curve can be explained by the fact that measuring errors may be about equal to, or even exceed, the real differences (differences corrected for measuring errors) between spectra of different compounds. The recall/reliability plots of Fig. 11 show the negative effect of missing peaks on the absolute behaviour of the similarity index: the plot for test set A indicates a lower reliability over the full range of the recall, as 27% of the correct positives are not retrievable, owing to the first preselection.

For a comparison of the C13RR system with the CIS/Clerc system, confusion plots for both systems, based on test set C, were generated (Fig. 12). These plots clearly demonstrate the better performance of the C13RR

TABLE 2

Results of a library search with the C13RR system and the CIS/Clerc system
[Test spectrum: "unknown" (one of two alternative library spectra of 1,3,3-trimethylbicyclo-2,2,1-heptan-2-one). Specified chemical shifts (ppm) were 222.3, 53.9, 47.2, 45.3, 41.6, 31.8, 24.9, 23.3, 21.6, 14.5. The trivial 100% matches are deleted]

Result of the C13RR system

In search for: unknown
Tolerance: 15.00 ppm (no missing peaks allowed)

No.	SI	Cas reg. nr.	Mol. formula/name
1.	38.2	1195795	C10 H16 01
	Bicyclo 2.2.1 heptan-2-one, 1,3,3-trimethyl-		
2.	8.5	76222	C10 H16 01
	Bicyclo 2.2.1 heptan-2-one, 1,7,7-trimethyl-		
3.	7.1	51100006	C10 H16 01
	Bicyclo 2.2.1 heptan-2-one 3,3,6-trimethyl-, *ENDO-*		
4.	5.3	56764320	C10 H16 01
	Bicyclo 3.2.1 Octan-6-one, 7,7-dimethyl-		
5.	4.2	4613461	C10 H16 01
	Bicyclo 2.2.1 heptan-2-one, 3.3.6-trimethyl, *EXO-*		

Result of the CIS/Clerc system

No.	Fit	Cas reg. nr.	Mol. formula/name
1.	97.29	41781-62-8	C16 H35 N1
	1-Hexanamine, *N*-butyl-*N*-(1,3-dimethylbutyl)-		
2.	97.12	20352-67-4	C8 H19 N1
	1-Hexanamine, *N*-ethyl-		
3.	96.85	4178-39-9	C12 H27 N1
	1-Hexanamine, *N*-(1,3-dimethylbutyl)-		
4.	96.68	1195-79-5	C10 H16 O1
	Bicyclo[2.2.1]heptan-2-one, 1,3,3-trimethyl-		
5.	96.24	123-82-0	C7 H17 N1
	2-Heptanamine		

system. As an illustration, the retrieval results obtained with both systems for an "unknown", which was one of two alternative library spectra for 1,3,3-trimethylbicyclo-2,2,1-heptan-2-one, are given in Table 2. It should be noted that the CIS/Clerc system includes the option of specifying multiplicities, in which case better results may be obtained.

Conclusions

An effective library search system should include a similarity index with a high discriminating power in the absolute as well as in the relative sense. The reproducibility based similarity index of the retrieval system for ^{13}C-n.m.r. spectra, reported in this paper, meets this requirement. The same concept of the similarity index has been used successfully in the development of a library search system for mass spectra. On the basis of the C13RR system, a microcomputer version (in UCSD Pascal) has also been developed. Results of this MC-C13RR system, as well as of the mass spectral reproducibility-based retrieval (MSRR) system, will be reported later.

The situation of "missing peaks" in unknown or reference spectrum causes a problem, which is dealt with in the present version of the library system by a simple modification of the search algorithm. Research aimed at further optimization of the retrieval systems, including a more adequate handling of missing peaks, is in progress.

The authors thank Dr M. J. A. de Bie for valuable discussions.

Appendix

Development of the reproducibility function. To rewrite Eqn. 6, the following function is introduced

$$p_0([\Delta q_i]/\sigma^2) = \int_{\Delta q^s=-\infty}^{\infty} p(\Delta q^s) \prod_{i=1}^{n} p(\Delta q_i^r)\, d\Delta q^s \qquad \text{(A1)}$$

This function is a normal probability function with non-zero correlations and can alternatively be expressed by

$$p_0([\Delta q_i]/\sigma^2) = A \exp(-0.5\,\chi^2) \qquad \text{(A2)}$$

where $A = (n+1)^{-1/2}\,(2\pi\sigma^2)^{-1/2n}$ and

$$\chi^2 = \sum_{i=1}^{n} (\Delta q_i - \overline{\Delta q^s})\Delta q_i/\sigma^2 \qquad \text{(A3)}$$

with

$$\overline{\Delta q^s} = \sum_{i=1}^{n} \Delta q_i/(n+1) \qquad \text{(A4)}$$

Here, χ^2 is a quantity distributed as chi-squared with n degrees of freedom. The quantity χ^2 can also be written as a function of S^2, being the estimator of σ^2

$$\chi^2 = n\,S^2/\sigma^2 \qquad \text{(A5)}$$

where

$$S^2 = \sum_{i=1}^{n+1} (\Delta q_i - \overline{\Delta q}^s)^2/n \tag{A6}$$

with $\Delta q_{n+1} = 0$. The $(n + 1)^{th}$ (artificial) difference in shift value can be considered as originating from the TMS peak in the reference and unknown spectra, and is, by definition, equal to zero. It may be noted that $\overline{\Delta q}^s$ can now also be expressed as a normal average value (cf. Eqn. A4): $\overline{\Delta q}^s = \Sigma_{i=1}^{n+1} \Delta q_i/(n + 1)$.

From Eqns. A1 and A2, the reproducibility function of Eqn. 6 can be rewritten as

$$p_0[\Delta q_i] = \int_{\sigma^2=0}^{\infty} p(\sigma^2)\, A \exp(-0.5\, \chi_n^2)\, d\sigma^2 \tag{A7}$$

According to Eqn. 4, the reproducibility function is characterised by the empirical parameters $E \ln \sigma^2$ and $V \ln \sigma^2$. From Eqn. A5 it follows that $\ln \sigma^2 = \ln S^2 - \ln(\chi^2/n)$. The expected value of $\ln(\chi^2/n)$ can be approximated [26] by $(-1/n - 1/3n^2)$. This implies that $E \ln \sigma^2$ can be estimated by averaging the quantity $\ln S^2 + 1/n + 1/3n^2$ for all difference spectra of the test set. If $E\sigma^2$ is defined as the expected value of σ^2, then from the assumption that σ^2 is lognormally distributed, it can be derived that

$$E\sigma^2 = \exp[E \ln \sigma^2 + 1/2\, V \ln \sigma^2] \tag{A8}$$

given that

$$V \ln \sigma^2 = 2[\ln(E\sigma^2) - E \ln \sigma^2] \tag{A9}$$

The value of $E\sigma^2$ can be approximated [25] by averaging S^2 for all difference spectra. The value of $V \ln \sigma^2$ can be found by substitution of the calculated values for $E\sigma^2$ and $E \ln \sigma^2$ in Eqn. A9.

Development of the similarity index. Elaboration of Eqn. 7 leads to

$$S_I = \int_{\sigma^2=0}^{\infty} p(\sigma^2)\, F(\sigma^2)\, d\sigma^2 \tag{A10}$$

with

$$F(\sigma^2) = \int_{\chi^2 < X^2} \ldots \int p_0([\Delta q_i] \mid \sigma^2)\, d\Delta q_1 \ldots d\Delta q_n \tag{A11}$$

where X^2 is the value of χ^2 (see Eqn. A3), calculated from the observed values of the difference quantities. This means that X^2 can be written as

$$X^2 = K/\sigma^2 \tag{A12}$$

where $K = \Sigma_{i=1}^{n} (\Delta Q_i - \overline{\Delta Q}^s)\, \Delta Q_i$ with $\overline{\Delta Q}^s = \Sigma_{i=1}^{n} \Delta Q_i/(n + 1)$.
Because χ^2 is distributed as chi-squared with n degrees of freedom, it follows from Eqn. A11 that

$$F(\sigma^2) = 1 - C(X^2) \tag{A13}$$

where C is the (cumulative) chi-squared distribution function with n degrees of freedom. Replacing, in Eqn. A10, the integration over σ^2 by an integration over $\ln \sigma^2$, and using Eqns. 4, A12 and A13 yield the following expression for S_I:

$$S_I = \int_{-\infty}^{+\infty} p(\ln \sigma^2)\, F'(\ln \sigma^2)\, d \ln \sigma^2 \tag{A14}$$

where $p(\ln \sigma^2) = N(\ln \sigma^2 : E \ln \sigma^2, V \ln \sigma^2)$ (A15)

and $F'(\ln \sigma^2) = 1 - C(K/e^{\ln \sigma^2})$ (A16)

This expression of S_I is the basis of the algorithm for the calculation of S_I values. The integration over $\ln \sigma^2$ (see Fig. 13) is done by numerical integration, approximating the functions $F'(\ln \sigma^2)$ and $p(\ln \sigma^2)$ by a fixed number of straight lines.

REFERENCES

1 W. Voelter, G. Haas and E. Breitmaier, Chem. Ztg., 97 (1973) 507.
2 P. R. Naegeli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
3 B. A. Jetzl and D. L. Dalrymple, Anal. Chem., 47 (1975) 203.
4 W. Bremser, M. Klier and E. Meyer, Org. Magn. Reson., 7 (1975) 97.
5 R. Schwarzenbach, J. Meili, H. Koenitzer and J. T. Clerc, Org. Magn. Reson., 8 (1976) 11.
6 J. Zupan, M. Penca, D. Hadzj and J. Marsel, Anal. Chem., 49 (1977) 2141.
7 D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, Org. Magn. Reson., 11 (1978) 535.
8 J. E. Dubois and J. C. Bonnet, Anal. Chim. Acta, 112 (1979) 245.
9 V. Mlynárik, M. Vida and V. Kellö, Anal. Chim. Acta, 122 (1980) 47.
10 W. Bremser, H. Wagner and B. Franke, Org. Magn. Reson., 15 (1981) 178.
11 M. Zippel, J. Mowitz, I. Kohler and H. J. Opferkuch, Anal. Chim. Acta, 140 (1982) 123.
12 A. P. Uthman, J. P. Koontz, J. Hinderliter-Smith, W. S. Woodward and C. N. Reilley, Anal. Chem., 54 (1982) 1772.
13 J. Kwiatkowski and W. Riepe, Anal. Chim. Acta, 135 (1982) 293.
14 P. Cleij, H. A. van 't Klooster and J. C. van Houwelingen, Anal. Chim. Acta, 150 (1983) 23.
15 E. Horowitz and S. Sahni, Fundamentals of Data Structures, Pitman Press, London, 1976.
16 L. F. Johnson and W. C. Jankowski, Carbon-13 NMR Spectra, Wiley-Interscience, New York, 1972.
17 F. I. Carroll, C. G. Moreland, G. A. Brine and J. A. Kepler, J. Org. Chem., 41 (1976) 996.
18 Y. Terui, K. Tori, S. Meada and Y. K. Sawa, Tetrahedron Lett., (1975) 2853.
19 C. J. Kelley, R. C. Harruff and M. Carmack, J. Org. Chem., 41 (1976) 449.
20 D. A. Forsyth, R. J. Spear and G. A. Olah, J. Am. Chem. Soc., 98 (1976) 2512.
21 H. Fritz and T. Winkler, Helv. Chim. Acta, 59 (1976) 903.
22 F. W. McLafferty, Anal. Chem., 49 (1977) 1442.
23 E. G. Smith, The Wiswesser Line—Formula Chemical Notation, McGraw-Hill, New York, 1968.
24 S. L. Grotch, Anal. Chem., 43 (1971) 1362.
25 M. Abramowitz and I. A. Segun (Eds.), Handbook of Mathematical Functions, Dover Publications, New York, 1968.