JOURNAL OF CHEMOMETRICS, VOL. 2, 11–27 (1988)

IDENTIFICATION AND SUBSTRUCTURE ANALYSIS OF OLIGOSACCHARIDE CHAINS DERIVED FROM GLYCOPROTEINS BY COMPUTER RETRIEVAL OF HIGH-RESOLUTION ^{1}H-NMR SPECTRA

D. S. M. BOT, P. CLEIJ AND H. A. VAN 'T KLOOSTER*

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

H. VAN HALBEEK,† G. A. VELDINK AND J. F. G. VLIEGENTHART

State University of Utrecht, Department of Bio-organic Chemistry, Padualaan 8, NL-3584 CH Utrecht, The Netherlands

SUMMARY

Based on a statistical model of the reproducibility of NMR spectral features, a system for computer retrieval of high-resolution ^{1}H-NMR spectra of glycoprotein carbohydrates has been developed. For corresponding peaks in an unknown and a reference spectrum, a similarity index based on the reproducibility of the chemical shifts is calculated. In addition, a second similarity index, based on the probability distribution of the percentage of non-matching peaks, has been developed. From these two similarity indices, a combined similarity index using the recall–reliability function as the optimizing criterion has been derived.

First results indicate that the '^{1}H-NMR reproducibility-based retrieval' ('1HRR') system offers good perspectives for both identification and substructure analysis.

KEY WORDS Oligosaccharides Glycoproteins Identification Substructure analysis High-resolution ^{1}H-NMR spectra Computer retrieval Reproducibility Chemical shifts Similarity index

INTRODUCTION

Various systems for the computer-aided library search of spectroscopic data have been developed.[1–16] These library search methods can be divided into two types: identification methods and interpretative methods. The main object of identification methods is to retrieve the reference data of the 'unknown' compound. In contrast, interpretative methods aim at retrieving data (if available) of compounds similar in structure to the unknown. A relatively large number of library search methods use mass, infrared or ^{13}C-NMR spectra or combinations thereof. Only a few systems have been developed for the computer-aided library

*Corresponding author; present address: National Institute of Public Health and Environmental Hygiene, Laboratory for Organic-analytical Chemistry, P.O. Box 1, 3720 BA Bilthoven, The Netherlands.

†Present address: Complex Carbohydrate Research Center, University of Georgia, PO Box 5677, Athens, Georgia 30613, U.S.A.

0886–9383/88/010011–17$08.50
© 1988 by John Wiley & Sons, Ltd.
Received 4 November 1986
Revised 9 April 1987

search of ^{1}H-NMR spectra.[13–16] In developing such systems, problems may arise because of the reproducibility of the spectra involved being dependent on a large number of external factors that have to be controlled carefully. Chemical shifts, linewidths and intensities suffer, *inter alia*, from considerable variations due to solvent, pH and temperature effects.

This paper reports the development and evaluation of a straightforward library search system for the identification of ^{1}H-NMR spectra of glycoprotein carbohydrates. The '1HRR' system uses chemical shifts (and no intensities or multiplicities) as the features to describe the spectrum. As the criterion of matching, a combined similarity index computed from two primary similarity indices is used. One of these similarity indices is based on a statistical model describing the reproducibility of chemical shifts and is applied to the matching peaks in spectra of the unknown and the reference compound. The other similarity index is based on the probability distribution of the percentage of mismatches between two spectra.

The general concepts for the development of these similarity indices were introduced by Cleij *et al.*[1] and elaborated for ^{13}C-NMR spectra by Bally *et al.*[2] An essential requirement for a straightforward library search method is that it should permit a ready classification of reference compounds into two groups: (i) compounds that could be and (ii) compounds that are very probably not identical to the unknown. In terms of hypothesis testing this is equivalent to establishing the truth or falsity of the null hypothesis that the unknown and the reference compound are identical. A library search system based on this principle should retrieve all references of the 'could be' class, i.e., all references with a similarity index exceeding a predefined threshold value, rather than five or ten 'best' matches (which may also be very bad matches).[1,2]

The main reasons for developing a retrieval system for identification of glycoprotein carbohydrates are the following:

(1) ^{1}H-NMR spectra of these compounds are highly informative on carbohydrate structure but fairly difficult to interpret;
(2) the rapid increase of the number of glycoprotein carbohydrates for which ^{1}H-NMR spectra have been recorded, and hence the availability of a considerable reference file (although for the development of the 1HRR system, only a limited database of some 60 spectra has been used).

It should be noted that ^{1}H-NMR spectra of (isolated) glycoprotein carbohydrates are usually recorded under well-defined constant conditions (solvent, pH and temperature), thereby circumventing the aforementioned irreproducibility problems.

REFERENCE DATA

In the library of carbohydrate reference ^{1}H-NMR spectra, only the chemical shift values of signals within certain ranges (1.0–3.5 and 4.0–5.6 ppm) are stored. These spectral regions contain more or less individually observable resonances ('structural reporter group signals').[17,18] According to the presence or absence of peaks in three prespecified ppm ranges (see below), the library is distributed over eight subfiles. Shift values are stored as output of the NMR computer; that is, having four decimal digits with a precision of 0·0001 ppm. The spectral information belonging to a reference spectrum is stored as a record in the library. Such a record contains the identifier of the spectrum, the chemical shifts belonging to the reporter group references and the compositional formula and molecular structure of the corresponding compound.

^{1}H-NMR spectra of glycoprotein carbohydrates can be quite complex, depending on the number of constituent monosaccharides, as illustrated by the example in Figure 1. This

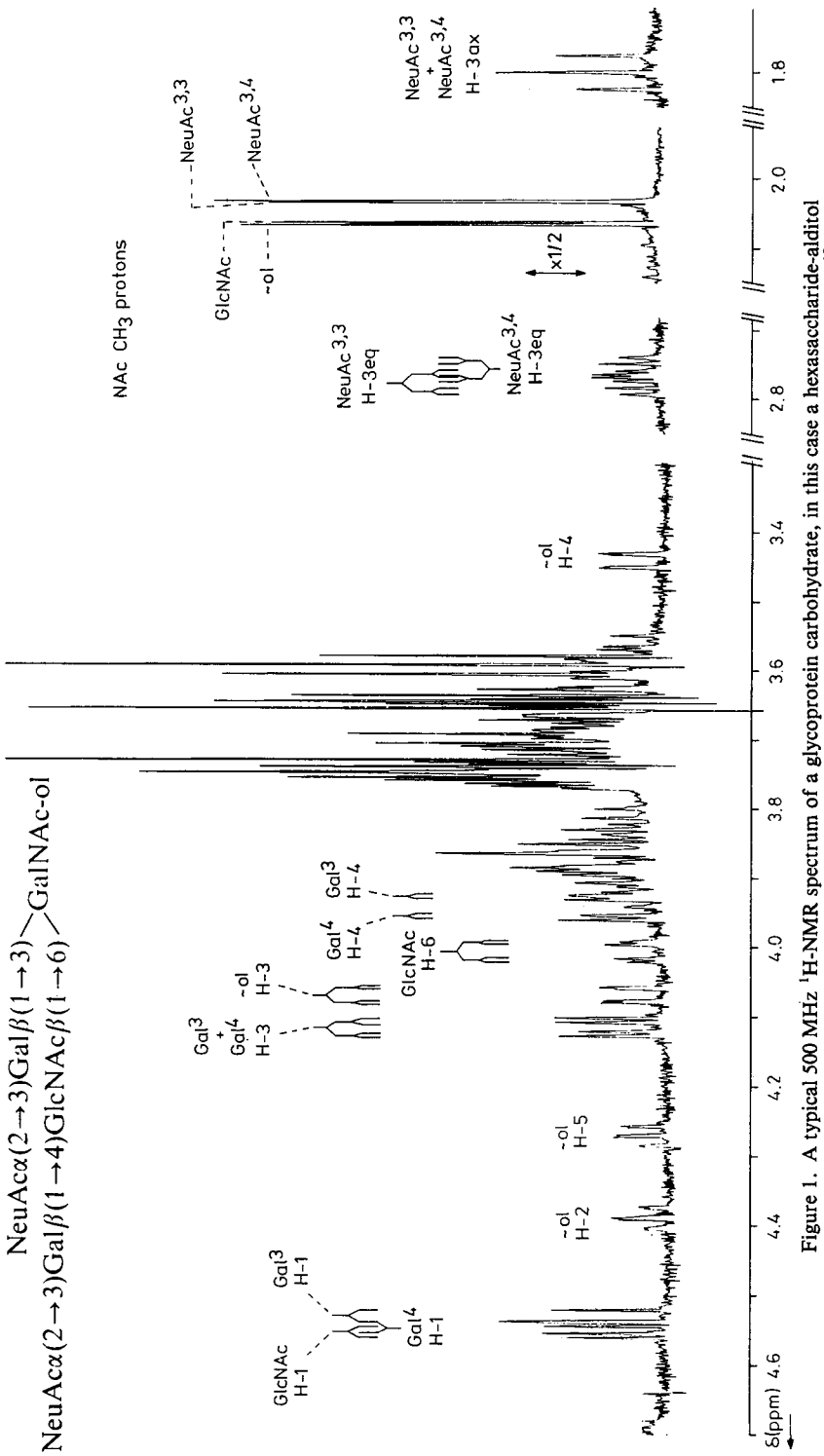

Figure 1. A typical 500 MHz ^{1}H-NMR spectrum of a glycoprotein carbohydrate, in this case a hexasaccharide-alditol originating from glycocalicin, the main part of a glycoprotein present in the human platelet membrane[19]

complexity is accompanied by a large amount of spectral information. However, not all of this information is needed for the retrieval of these spectra. To obtain data reduction the following restrictions are made:

(1) Chemical shift values within the range 3.5–4.0 are not considered by the system because of their low directly accessible information content, due to partial overlap of peaks in this spectral region.
(2) Multiplicities are not used because this information is not available *a priori*. (The 1HRR system is developed for handling raw spectral data; i.e., all the spectral information within the specified ranges that can be obtained directly from a spectrum without the need for interpretation by a spectroscopist.
(3) Relative intensities are not used because of the poor reproducibility of NMR spectral integration techniques.

The remaining spectral information, consisting of chemical shift values of each peak of the multiplets within the predefined ppm ranges (1.0–3.5 and 4.0–5.6 ppm), in practice provides enough information for the correct retrieval of reference spectra.

COMPARISON OF SPECTRA. GENERAL STRATEGY

The peaks in the preselected regions of an unknown and a reference spectrum generally can be divided into a set of (pairs of) corresponding ('matching') peaks and a set of peaks without a counterpart in the other spectrum ('mismatching' peaks). The number of mismatching peaks may be non-zero, even if the unknown and reference spectra originate from identical compounds. This effect may be due to variations in the experimental conditions (field strength, temperature, pH, sample concentration) under which the ^{1}H-NMR spectra are recorded. A distinct small peak in one spectrum of some compound may 'vanish' in the noise of another spectrum (recorded under different experimental conditions) of the same compound. Also, peaks that are separated in a high-field spectrum may coincide with another, lower-field, spectrum of the same compound. In other words, a retrieval algorithm for ^{1}H-NMR spectra should take into account mismatching peaks and the fact that spectra of identical compounds may show a different number of peaks.

To evaluate the similarity between unknown and reference spectra, both the shift differences for matching peaks and the (number of) mismatching peaks have to be taken into account; these parameters are to be combined in a single evaluation criterion measuring the overall similarity of both spectra.

To define exactly what are matching and what are non-matching peaks, the window concept is introduced. The window is defined as the maximum distance by which peaks in the spectra compared may be separated, yet should be considered as corresponding peaks. The window should be wide enough to make sure that nearly all shift differences, observed for corresponding peaks in alternative spectra of the same compound, will be smaller than or equal to the applied window. On the other hand, there are reasons for not choosing the window too wide. If, for example, the spectra compared contain 60 and 45 peaks, respectively, and the window applied is very wide, then there are $\binom{60}{45} = 5{\cdot}32 \times 10^{13}$ possibilities to match these spectra. By reducing the window width, the number of possibilities and thus the computer consumption time can be reduced drastically. Another reason for not choosing too wide a window is that non-corresponding peaks should not be 'matched' with each other, because this would result in an incorrect value of the similarity index. The optimal window width for the 1HRR system appeared to be 0·0040 ppm. As can be seen from Figure 2, the window width is larger than most shift differences observed.

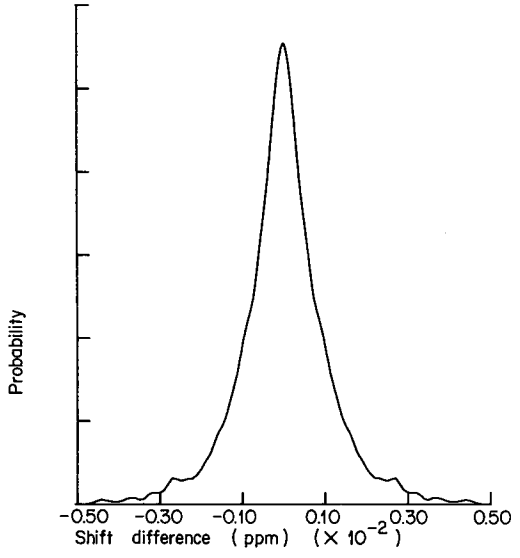

Figure 2. Probability distribution of difference in shift values derived from 61 pairs of 'alternative spectra': spectra of the same compound recorded under highly different experimental conditions

However, an ambiguous situation may still occur in which a peak in one spectrum potentially matches with two or more peaks in the other spectrum. In order to avoid such ambiguity, a further requirement is introduced here, expressing that the set of matching peaks should be chosen such that, considering the reproducibility, it constitutes the most probable configuration of corresponding peaks. This is further elaborated in the next section.

A SIMILARITY INDEX FOR MATCHING PEAKS

The criterion for the evaluation of the differences in shift values for matching peaks is based on the similarity index (SI) proposed by Cleij *et al.*[1] It implies the use of so-called difference quantities, representing for two spectra the differences in value of a set of feature quantities selected to describe a spectrum. The use of chemical shifts of a set of n matching peaks 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 pair of matching peaks and n is the number of matching peak pairs.

The similarity index SI is defined in terms of a model of the reproducibility of the feature quantities or, more precisely, in terms of the reproducibility function $p_0[\Delta q_i]$ measuring the probability of observing a combination of differences $\Delta q_1, \Delta q_2, \ldots, \Delta q_n$ for a difference spectrum derived from two alternative spectra.

For the development of the reproducibility model of chemical shifts, a set of pairs of alternative spectra (spectra of the same compound recorded under different experimental conditions, one spectrum to be considered as a replicate of the other) was used. In designing the model of reproducibility, it is assumed that these alternative spectra are representative of the unknown and target reference spectra occurring in actual search situations. These pairs of alternative spectra were analysed in the form of 'difference spectra'. A difference spectrum of

two spectra A and B is defined as a plot of the differences in shift value for corresponding peaks versus the shift values of these peaks of spectrum A.

Inspection of some 60 difference spectra showed that the same model used by Bally *et al.*[2] for the 13CRR system is applicable to the retrieval of ^{1}H-NMR spectra. This implies that the differences in shifts can be thought of as consisting of a random and a systematic part.

As illustrated by Figure 3, the random part is indicated by the scatter of the data points in a difference spectrum around the average value (broken line). The systematic part of the differences in shift values is represented by the deviation of this average difference from zero. Such systematic deviations will result from variations in the position of the chemical shift calibration peak (usually, internal acetone at $\delta 2.225$). In other words, the total difference in a chemical shift for peak i, Δq_i, consists of a random part Δq_i^r and a systematic part Δq^s, i.e.,

$$\Delta q_i = \Delta q_i^r + \Delta q^s \quad (\text{for } i = 1 \ldots n) \tag{2}$$

Further assumptions are that the statistical variations in peak position (including the calibration peak) can, within a single difference spectrum, be described by a normal distribution with variance σ^2. In order to allow statistical variations in the value of σ^2 for the various difference spectra, σ^2 is considered as a stochastic quantity obeying some probability distribution. As in the MSRR system for mass spectra[1] and the C13RR system for ^{13}C-NMR spectra,[2] this function is approximated here by a log-normal distribution function. According to this model, the reproducibility function $p_0[\Delta q_i]$ can be written as

$$p_0[\Delta q_i] = \int_{\sigma^2=0}^{\infty} \mathrm{LN}(\sigma^2: E \ln \sigma^2, V \ln \sigma^2) \int_{\Delta q^s=-\infty}^{\infty} N(\Delta q^s: 0, \sigma^2) \prod_{i=1}^{n} N(\Delta q_i^r: \Delta q^s, \sigma^2)\, \mathrm{d}\Delta q^s\, \mathrm{d}\sigma^2 \tag{3}$$

where $\mathrm{LN}(x: \mathrm{EL}, \mathrm{VL})$ is a log-normal probability function for variable x, with EL and VL being the expected value and variance of $\ln x$ respectively; $N(x: E, V)$ is a normal probability function for variable x, with E and V being the expected value and variance of x respectively. This model contains two empirical parameters, $E \ln \sigma^2$ and $V \ln \sigma^2$, that can be calculated from some representative set of difference spectra derived from pairs of alternative ^{1}H-NMR spectra.

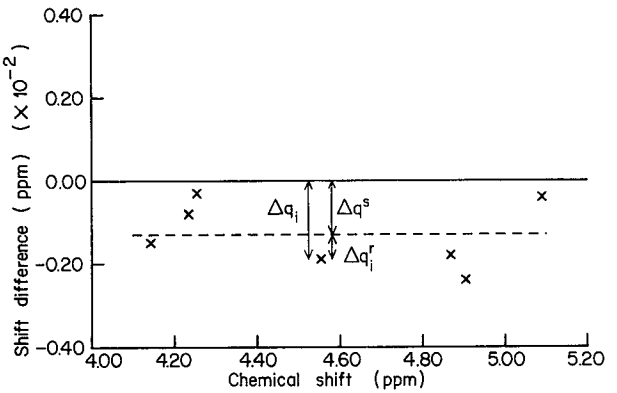

Figure 3. A difference spectrum of two spectra A and B is a plot of the difference in shift values for corresponding peaks versus the shift values of the matching peaks in spectrum A. The differences consist of a random part (Δq_i^r) and a systematic part (Δq^s), in this case amounting to $-0{\cdot}0007$ and $-0{\cdot}0013$ ppm respectively

The general form of the similarity index SI is given by

$$SI = \int_{\substack{\Delta q_1 \ldots \int_{\Delta q_n} \\ R[\Delta Q_i]}} d\,\Delta q_1 \ldots d\,\Delta q_n \tag{4}$$

where $R[\Delta Q_i]$ is the region in the multidimensional 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.[1]

Elaboration of this expression using the reproducibility function described above leads to the following expression for SI_{shifts}, the similarity index applied to chemical shift difference:

$$SI_{shifts} = \int_{\ln \sigma^2 = -\infty}^{\infty} N(\ln \sigma^2 : E \ln \sigma^2,\ V \ln \sigma^2)[1 - C(K/e^{\ln \sigma} : n)]\ d \ln \sigma^2 \tag{5}$$

where $C(x: N)$ is the (cumulative) chi-squared distribution function for variable x with N degrees of freedom and K is given by

$$K = \Sigma(\Delta q_i^2) - (\Sigma\, \Delta q_i)^2/(n+1) \tag{6}$$

In the 1HRR system the integral over $\ln \sigma^2$ is calculated by a special form of numerical integration using tabulated values of the chi-squared function.[20]

As for the requirement that a chosen set of matching peaks should represent the most probable configuration, it can now be shown that the quantity K is directly related to the probability (derived from the reproducibility function) of observing certain combinations of shift differences for a set of corresponding peak pairs, as found for alternative spectra. The lower this probability for a set of peak pairs is, the lower the probability that this set involves a set of corresponding peaks. Hence the set of matching peaks can be found by minimizing the calculated K-value (this corresponds to maximizing the SI-value) of the set.

The following algorithm was used to find this set. The starting point is that two successive peaks in a spectrum, having chemical shift values that differ by more than 2 × window width, can never match with the same peak in another spectrum. These data are used to divide the

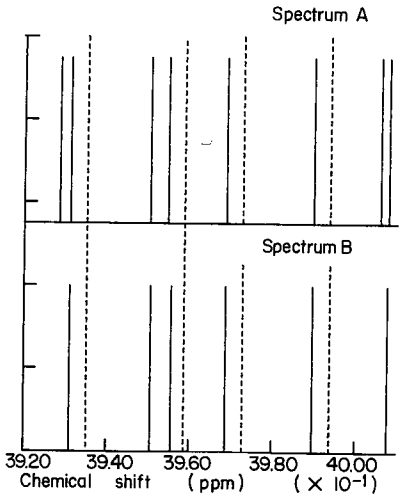

Figure 4. Apart from deviations in corresponding shift values, alternative spectra show different numbers of peaks. If a spectrum A containing 60 peaks is compared with a spectrum B with 45 peaks, there are $5{\cdot}32 \times 10^{13}$ possibilities to delete 15 peaks. A reduction is achieved by dividing the spectrum into subspectra (separated by boken lines)

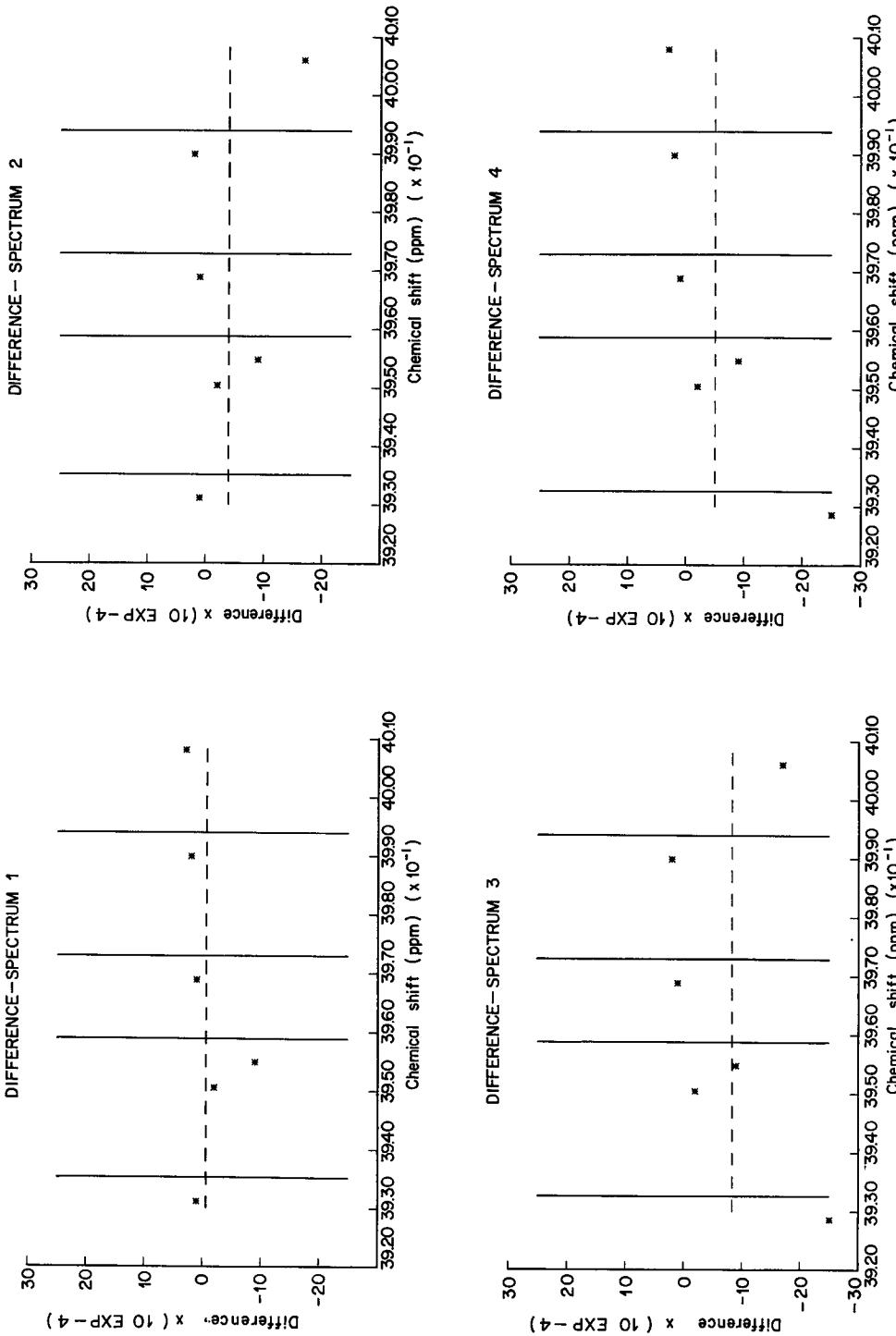

Figure 5. For each combination of subspectra, one or more difference subspectra can be calculated. From the optimal combination of difference subspectra, the entire difference spectrum is calculated

spectra compared into subspectra (Figure 4). For each pair of subspectra being compared, all possible difference subspectra are computed (Figure 5).

The only information that is needed from such a difference subspectrum are the values of $\Sigma\,\Delta(q_i)$ and $\Sigma\,\Delta(q_i^2)$. For all the difference subspectra these values are stored. When the complete spectra are compared, an optimal combination of the difference subspectra is obtained by calculating the minimum value for the K-parameter. By recursively scanning all possible combinations of difference subspectra, the optimal combination can be computed.

Another problem arises because of the window being applied. When there is a systematic deviation present in the (total) difference spectrum, a wrong value for SI_{shifts} and $SI_{mismatch}$ may be computed. This problem can be illustrated graphically as in Figure 6.

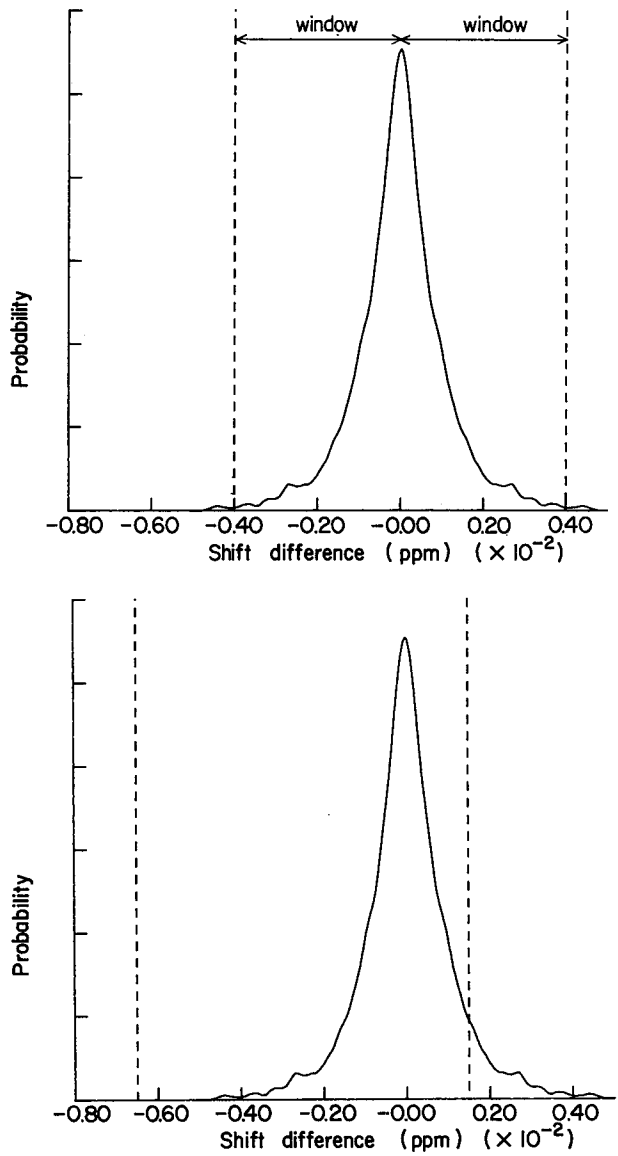

Figure 6. A systematic deviation of −0·0025 ppm introduces ~10% extra mismatches

Because of systematic deviations, an extra number of mismatches can be expected and therefore a wrong value for both SI_{shifts} and the number of mismatches might be computed. The solution to this problem is to require that the systematic deviation will not exceed a certain value (0·0008 ppm). When a systematic deviation is computed greater than this value, all the shifts in one of the two spectra, except for the shift corresponding to the calibration peak, are corrected for this deviation. The reason for doing this is that systematic deviations are in fact reflections of variations in the absolute chemical shift values of calibration peaks.

Using the above elaborated strategy of finding the corresponding peaks in two alternative spectra, the values of $E \ln \sigma^2$ and $V \ln \sigma^2$ were calculated from 61 pairs of alternative spectra, resulting in the values 4.614 and 0.3094 respectively.

A SIMILARITY INDEX BASED ON THE MISMATCH PERCENTAGE

A mismatch occurs when, for a certain peak in spectrum A, there is no corresponding peak in spectrum B, or vice versa. For two spectra A and B the mismatch percentage (MP) is defined as follows:

$$MP = \frac{\text{total number of mismatches}}{\text{sum of the numbers of peaks of spectrum A and B}} \tag{7}$$

Given the fact that SI_{shifts} has the form of a *P*-value as used in hypothesis testing,[1] this form should also be chosen for a matching criterion based on the MP, in order to enable the formulation of a single evaluation criterion for both shift differences and number of mismatches. If, for the comparison of unknown and reference spectra, the MP is considered to be used as the test quantity, then it is necessary to establish the probability distribution of MP for difference spectra of pairs of alternative spectra, and the same distribution also for the case that the unknown and reference compounds are different (Figure 7).

In testing the null hypothesis using a *P*-value, the null hypothesis is considered to be rejected (unknown and reference compound not identical) at a significance level α, when $P < \alpha$. A test with maximum power is constructed when the *P*-value is obtained by integrating the probability

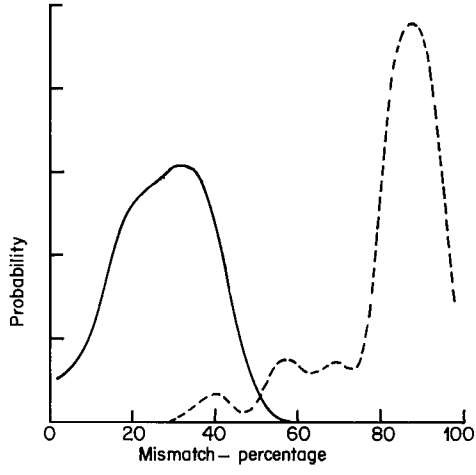

Figure 7. Probability densities of the percentage of missing peaks (MP) have been determined for the set of 61 difference spectra of identical compounds (——) and for a set of difference spectra of non-identical compounds (---). The separation of these curves shows why the MP parameter can be used as the basis of a second similarity index ($SI_{mismatch}$), which also has the form of a significance probability (*P*-value)

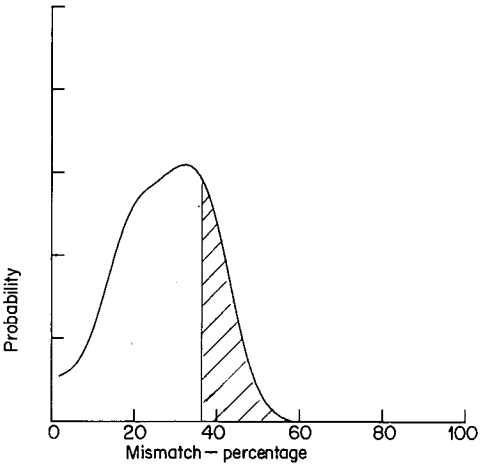

Figure 8. Calculation of the similarity index based on the percentage of missing peaks ($SI_{mismatch}$ = shaded area)

function of MP for alternative spectra over a region defined by

$$\frac{p_0(\mathrm{MP})}{p_1(\mathrm{MP})} \leqslant \frac{p_0(\mathrm{MP_{obs}})}{p_1(\mathrm{MP_{obs}})} \tag{8}$$

where p_0 is the probability density of MP under the null hypothesis, p_1 is the probability density under the alternative hypothesis and MP_{obs} in the observed value of MP. This condition can be reduced to $MP \leqslant MP_{obs}$, assuming that $p_0(MP)/p_1(MP)$ is a monotonically decreasing function of MP. In other words, the assumption is made that for $MP_{obs} = 0$ it is most probable that reference and unknown compounds are identical, and that the more MP is deviating from zero the more probable it is that unknown and reference compounds are different. Therefore we define $SI_{mismatch}$ as

$$SI_{mismatch} = \int_{MP=MP_{obs}}^{MP=100} p_0(MP)\ dMP \tag{9}$$

$SI_{mismatch}$ values are calculated from the empirical distribution of MP determined from 61 pairs of alternative spectra (see Figure 8).

A COMBINED SIMILARITY INDEX

For the determination of the optimal way of combining SI_{shifts} and $SI_{mismatch}$, recall–reliability plots[3] were used as the evaluation criterion.

The recall for a test set of 'unknown' spectra is defined as the number of target reference spectra actually retrieved, divided by the total number of target spectra available in the test set. The reliability is defined as the number of retrieved target spectra, divided by the total number of retrieved reference spectra.

Recall–reliability plots were obtained by using a test set of 61 spectra, each of which had a replicate (alternative) spectrum in the reference database. SI_{comb} was computed from SI_{shifts} and $SI_{mismatch}$ by addition and normalization of the sum. Further optimization was required to obtain the optimal weight factor (w_{sm}):

$$SI_{comb} = (w_{sm}SI_{shifts} + SI_{mismatch})/(w_{sm} + 1) \tag{10}$$

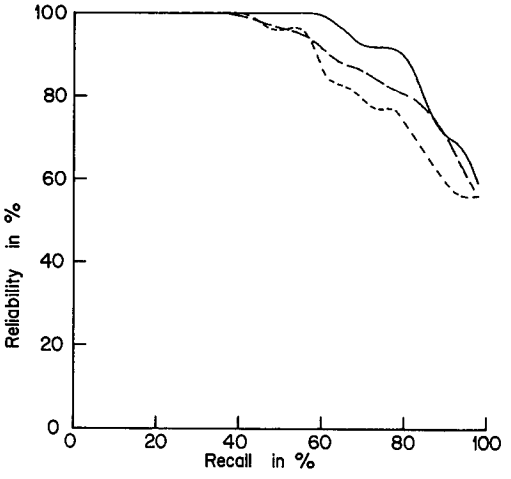

Figure 9. The similarity indices SI_{shifts} and $SI_{mismatch}$ are combined into SI_{comb}, using the recall–reliability function as the optimization criterion, resulting in

$$SI_{comb} = (2 \cdot 0SI_{shifts} + SI_{mismatch})/3 \cdot 0$$

It is shown that SI_{comb} (——) provides better retrieval results than each of SI_{shifts}(– –) and $SI_{mismatch}$(---) separately

The outcome of this optimization was that the best retrieval results can be expected using a weight factor with a value close to 2. Figure 9 shows that combining SI_{shifts} and $SI_{mismatch}$ provides a much better performance for the retrieval system than could have been achieved by using only one of these. The results depicted in Figure 9 were obtained using preselection criteria, as described in the next section.

PRESELECTIONS

The following preselections are applied within the 1HRR system:

(1) On the basis of the presence or absence of peaks in three prespecified regions (5·60–4·77,

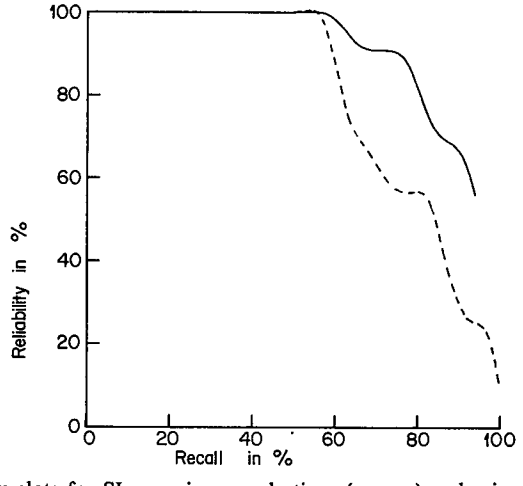

Figure 10. Recall–reliability plots for SI_{comb} using preselections (———) and using no preselections (— — —)

1·95–1·60, 1·30–1·00), each indicating the presence of a particular structure element, one of the eight subfiles is selected for retrieval.

(2) Reference spectra for which a $SI_{mismatch}$ value less than 2% is calculated are not further processed.

As can be seen from Figure 10, the use of preselections generally improves the recall–reliability behaviour of the system.

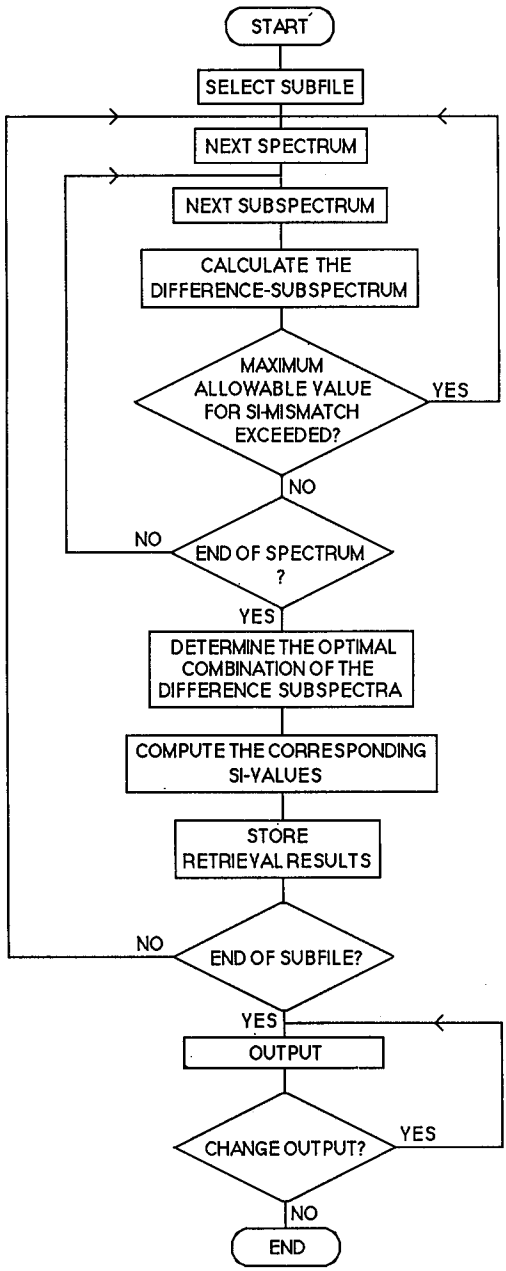

Figure 11. Flow diagram of the retrieval algorithm (main steps)

UNIVERSITY OF UTRECHT
HF-1H-NMR RETRIEVAL SYSTEM
DATABASE: CARBOHYDRATES

(a) UNKNOWN : TESTSAMPLE 1, pre-identified as: NEUACa(2-3)GALb(1-3)
GANOL
NEUACa(2-3)GALb(1-4)GLNb(1-6)

THRESHOLD SI_{shifts} = 0 %
THRESHOLD $SI_{mismatch}$ = 2 %

RETRIEVAL RESULTS:

NR	IDENTIF	SI_{comb} (%)	SI_{shifts} (%)	SI_{mism} (%)	STRUCTURE
1)	AP.001	54	53	56	NEUACa(2-3) GALb(1-3) GANOL NEUACa(2-3)GALb(1-4)GLNb(1-6)
2)	AP.002	53	44	71	NEUACa(2-3) GALb(1-3) GANOL NEUACa(2-3)GALb(1-4)GLNb(1-6)
3)	AP.003	51	34	86	NEUACa(2-3) GALb(1-3) GANOL NEUACa(2-3)GALb(1-4)GLNb(1-6)
4)	AP.004	23	7	54	NEUACa(2-3) GALb(1-3) GANOL GALb(1-4)GLNb(1-6)
5)	BH.001	14	18	8	NEUACa(2-3)GALb(1-3) GANOL
6)	BJ.002	13	18	3	NEUACa(2-3)GALb(1-3) GANOL
7)	CJ.001	12	11	13	NEUACa(2-3)GALb(1-3) GANOL GALb(1-4)GLNb(1-6)
8)	DP.001	10	11	8	NEUACa(2-3)GALb(1-3) GANOL
9)	EJ.001	9	7	14	NEUACa(2-3)GALb(1-3) GANOL GALb(1-4)GLNb(1-6)
10)	FJ.001	5	3	10	NEUACa(2-3)GALb(1-3) GANOL NEUACa(2-6)

END OF LIST

Figure 12. Typical outputs of the 1HRR system modules. The system starts in a CONTROL module, which allows the user 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. The module RETRIEVE starts the library search for a specified spectrum and presents the result of the search in the form of a list of references having a specified minimum similarity to the unknown. In this case (Figure 12(a)), for a spectrum of a known test compound, consisting of two chains attached to a GaINAc-ol unit, the system has found three target reference spectra present in the database ('correct positives', ranked as numbers 1, 2 and 3). The reference compounds retrieved with numbers 4–10 show lower SI values, but are all structurally related to the target reference compound. Declaration of codes used:

GALb(1–3) = Galβ(1–3)
GALb(1–4) = Galβ(1–4)
GANol = GalNAc-ol
GLNb(1–6) = GalNAcβ(1–6)
NEUACa(2–3) = NeuAcα(2–3)
NEUACa(2–6) = NeuAcα(2–6)

```
**********************************************************************************************

(b)  >> SEARCH INITIATED <<

SELECT THE OPTION YOU WANT TO USE; TYPE H(ELP) FOR INFORMATION :  H
THE FOLLOWING OPTIONS ARE AT YOUR DISPOSAL:
   B = EXACT MATCH OF COMPOSITIONAL FORMULA
   S = SEARCH ON SHIFT INTERVAL
   I = SEARCH ON IDENTIFIER
   E = END

SELECT THE OPTION YOU WANT TO USE; TYPE H(ELP) FOR  INFORMATION:  S

HOW MANY INTERVALS DO YOU WANT TO USE FOR MATCHING  (MAX=5) ?      2
ENTER THE START VALUE OF INTERVAL NUMBER 1:  52200
ENTER THE END VALUE OF INTERVAL NUMBER 1   :  52150
ENTER THE START VALUE OF INTERVAL NUMBER 2:  20600
ENTER THE END VALUE OF INTERVAL NUMBER 2   :  20500

DO YOU WANT THE STRUCTURES TO BE  DISPLAYED (Y/N) ?       Y
DO YOU WANT THE CHEMICAL SHIFTS TO BE  DISPLAYED (Y/N) ?  Y
DO YOU WANT THE LISTING TO BE SAVED FOR  PRINTING (Y/N) ?  N

SEARCHING FOR:  SPECTRA WITH CHEMICAL SHIFTS
BETWEEN:   52200  AND  52150
BETWEEN:   20600  AND  20500

THE FOLLOWING REFERENCE COMPOUNDS ARE FOUND:
----------------------------------------------------------------------------------------------

SL9.01

NEUACa (2-6 )GALb (1-4) GLNb (1-2) MANa (1-3)
                                          MANb (1-4) GLNa/b
NEUACa (2-6) GALb (1-4) GLNb (1-2) MANa (1-6)

CHEMICAL SHIFTS:
52182  52118  51366  49506  47844  47755  46148  45990  44527  44369
42628  42562  42516  42000  41961  41246  41220  41182  41156  40080
39899  39689  39548  39505  39312  39246  39169  39114  39037  26866
26795  26773  26618  26531  22011  20699  20663  20631  20603  20517
20495  20304  17437  17193  16951
----------------------------------------------------------------------------------------------

MA4.01

NEUACa (2-6 )GALb (1-4) GLNb (1-2) MANa (1-3)
                                          MANb (1-4) GLNa/b
NEUACa (2-6) GALb (1-4) GLNb (1-2) MANa (1-6)

CHEMICAL SHIFTS:
52181  52117  51358  49518  49442  46146  45989  44522  44364  42653
42576  42541  42013  41975  41261  41232  41194  41169  40077  39897
39688  39557  39507  39311  39260  39181  39126  39068  39018  26859
26792  26766  26611  26523  20810  20705  20689  20664  20632  20601
20528  20488  20297  20081  17426  17182  16938
----------------------------------------------------------------------------------------------

SELECT THE OPTION YOU WANT TO USE; TYPE H(ELP) FOR  INFORMATION :  E
DO YOU WANT A PRINTED LISTING OF THE SEARCH RESULTS (Y/N) ?        N

>> END SEARCH <<
**********************************************************************************************
```

The module FIND allows the entering of a search key (spectrum identifier, compositional formula or chemical shift intervals), which results in a list of reference compounds containing the search key. In Figure 12(b) the results are shown for a search key consisting of two shift intervals. The module ADDTOLIB enables the user to add new spectra with identifying information to the library

IMPLEMENTATION OF THE 1HRR SYSTEM

For the design, development and evaluation of the 1HRR system, a Data General Eclipse MV/4000 computer in our laboratory was used. Programs were written in Pascal. The spectra stored in the library were all recorded on a Bruker 500 MHz ^{1}H-NMR spectrometer (WM500), located at the SON Dutch NMR facility, Department of Biophysics, University of Nijmegen, The Netherlands.

The main steps of the retrieval algorithm can be illustrated by the flow diagram in Figure 11. The 1HRR system consists of four modules. The program starts in the module CONTROL, which contains the system commands and controls the input and storage of the unknown spectra. The module FIND enables the user to select spectra from the reference files by specifying features belonging to the spectrum. The module ADDTOLIB can be used to store new spectra with their spectral information in the library. Typical outputs of some modules are given in Figure 12.

RESULTS AND DISCUSSION

The recall–reliability plots of Figures 9 and 10 show that the results of the 1HRR system for the database used are not only encouraging but also indicate that for larger databases acceptable results can be expected. The output shown in Figure 12(a) reveals that the combination of handling subspectra and allowing different numbers of peaks in the comparison of spectra results in the identification of structurally related molecules, the smaller molecules being substructures of the larger ones. If, for example, the hit list contains a reference spectrum with a low value of $SI_{mismatch}$ and a high value of SI_{shifts}, the compound represented by this reference spectrum probably contains the same substructure as the compound represented by the unknown spectrum. The 1HRR system allows the user to reject the second preselection, so that reference spectra from all subfiles are compared with the unknown spectrum. The user can choose this option if in the preselected subfile no relevant information is found. By searching the whole library, spectra containing similar substructures may be found.

CONCLUSIONS

The combination of two reproducibility-based similarity indices has resulted in an effective matching criterion for retrieval of high-resolution ^{1}H-NMR spectra of glycoprotein carbohydrates. It is shown that the chemical shifts within the ranges 1·0–3·5 ppm and 4·0–5·6 ppm contain sufficient information to enable identification and substructure analysis; i.e., if the relevant (sub)structures are contained in the database. Peak multiplicities can thus be ignored.

Although the relatively small library used for this investigation does not allow a comprehensive evaluation nor a generalization of the conclusions, the first results of the 1HRR system are quite promising.

ACKNOWLEDGEMENTS

The authors wish to thank Dr J. H. G. M. Mutsaers for her assistance in recording part of the NMR spectra, and Dr M. van Iwaarden for his valuable contributions during the development of the 1HRR system. This investigation was supported in part by the Netherlands Foundation for Chemical Research (SON/ZWO) and the Netherlands Foundation for Cancer Research (KWF, grant UUKC-OC79-13).

REFERENCES

1. P. Cleij, H. A. van 't Klooster and J. C. van Houwelingen, *Anal. Chim. Acta* **150**, 23 (1983).
2. R. W. Bally, D. van Krimpen, P. Cleij and H. A. van 't Klooster, *Anal. Chim. Acta* **157**, 227 (1984).
3. F. W. McLafferty, *Anal. Chem.* **49**, 1442 (1977).
4. W. Voelter, G. Haas and E. Breitmaier, *Chem. Ztg.* **97**, 507 (1973).
5. P. R. Naegeli and J. T. Clerc, *Anal. Chem.* **46**, 739 (1974).
6. R. Schwarzenbach, J. Meili, H. Koenitzer and J. T. Clerc, *Org. Magn. Reson.* **8**, 11 (1976).
7. J. Zupan, M. Penca, D. Hadzj and J. Marsel, *Anal. Chem.* **49**, 2142 (1977).
8. D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, *Org. Magn. Reson.* **11**, 535 (1978).
9. H. B. Woodruff, C. R. Snelling, Jr., C. A. Shelley and M. E. Munk, *Anal. Chem.* **49**, 2075 (1977).
10. W. Bremser, H. Wagner and B. Franke, *Org. Magn. Reson.* **15**, 178 (1981).
11. A. P. Uthman, J. P. Koontz, J. Hinderliter-Smith, W. S. Woodward and C. N. Reilley, *Anal. Chem.* **54**, 1772 (1982).
12. J. Kwiatowski and W. Riepe, *Anal. Chim. Acta* **135**, 293 (1982).
13. V. Mlynarik, M. Vida and V. Kello, *Anal. Chim. Acta* **122**, 47 (1980).
14. Y. Katagiri, K. Kanohta, K. Nagasawa, T. Okusa, T. Sakai, O Tsumura and Y. Yotsui, *Anal. Chim. Acta* **133**, 535 (1981).
15. E. F. Hounsell, D. J. Wright, A. S. R. Donald and J. Feeney, *Biochem. J.* **223**, 129 (1984).
16. D. R. Anderson and W. J. Grimes, *Anal. Biochem.* **146**, 13 (1985).
17. J. F. G. Vliegenthart, L. Dorland and H. van Halbeek, *Adv. Carbohyd. Chem. Biochem.* **41**, 209 (1983).
18. K. Bock and H. Thøgersen, *Ann. Rep. NMR Spectrosc.* **13**, 1 (1982).
19. S. A. M. Korrel, K. J. Clementson, H. van Halbeek, J. P. Kamerling, J. J. Sixma and J. F. G. Vliegenthart, *Eur. J. Biochem.* **140**, 571 (1984).
20. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, p. 984, Dover Publications, New York (1968).