CASPER: A Computer Program Used for Structural Analysis of Carbohydrates

PER-ERIK JANSSON, LENNART KENNE, and GÖRAN WIDMALM*

Department of Organic Chemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

Received February 18, 1991

The computer program CASPER and its algorithms are described. The program is aimed at facilitating the determination of structures of oligosaccharides and regular polysaccharides, requiring as input either the one-dimensional ^{1}H or ^{13}C NMR spectrum or the 2D C,H-correlation NMR spectrum together with information on components and linkages. The databases, the method of simulating spectra, options of the program, and techniques for faster calculations are described as well as an example of a structural determination.

INTRODUCTION

In the realm of natural products, carbohydrates constitute an important class. In plants, they may act as storage materials or as structural components. In bacteria, they surround the cell as a protective shield against the defense system of the host, and they also express the immune specificity. In higher organisms, glycoconjugates such as glycoproteins and glycolipids are involved in several recognition phenomena, e.g., in cell–cell recognition. The blood-group specificity in man is determined by different oligosaccharide elements. A large number of monosaccharides have been identified in nature, and several combinations of these are possible. This makes the diversity of carbohydrate structures large, which gives nature a way of expressing various messages through different structures.

In order to understand properly interactions at the molecular level, knowledge of the primary structure of carbohydrates (see below) and their solution conformation together with the molecular dynamics is necessary. The determination of these features is often a difficult task and has been hampered by the fact that determination of the primary structure is rather time-consuming. This has prompted faster and simpler methods of structure determination.

Polysaccharides can be divided into irregular and regular ones. The irregular polysaccharides are found mainly in plants, while the regular ones mostly have microbial origin and are repetitive sequences of oligosaccharide units. The reason for the latter is that the polymer is synthesized from oligosaccharide–polyprenol conjugates. Determination of the primary structure of oligosaccharides and regular polysaccharides includes the determination of the following: the constituent sugars and their anomeric configuration, the linkage position(s), the sequence of the sugars and if present, substituents and their location. The sequence determination is usually the most time-consuming part. Previously, identification of fragments obtained from specific degradations, e.g., enzymatic digestions or specific chemical degradations were commonly used to deduce the sequence. Today NMR spectroscopy is used frequently, but in order to determine partial or whole sequences of a structure NMR spectra have to be assigned. The assignment may be both difficult and time-consuming; therefore a method in which assignments are not necessary should greatly simplify the determination.

Automated approaches for the structural elucidation of natural products using computer programs have appeared.[1,2] For oligosaccharides[3-5] and polysaccharides[6,7] different programs using NMR data have been developed. We have previously shown that by using the computer program CASPER (Computer Assisted SPectrum Evaluation of Regular polysaccharides) it is possible to deduce the primary structure of oligosaccharides[8] and polysaccharides with repeating units.[9,10] We now report the details and algorithms of the program.

OVERVIEW

The menu hierarchy in CASPER is shown in Scheme I. Most of the menus will be discussed in detail below. Information concerning all details about CASPER can be obtained under the menu 'Information'. Structural analysis of an oligo- or polysaccharide can be performed under the menu 'Structural Determination'. The generalized structural determination flow scheme is shown in Scheme II.

In brief, the CASPER program works as follows: given data on sugar components and the linkages obtained from sugar and methylation analysis, the program calculates all possible structures and simulates their ^{1}H and ^{13}C NMR spectra. This is done by an additive approach in which the chemical shifts of the monosaccharides and chemical shift displacements, different for different types of glycosidic linkages, are summed. Below, the chemical shift displacements are called the glycosylation shifts. To a first approximation only short-range interactions, between two neighboring sugars, are assumed to be present. This in not always true for branch points and modifications must be introduced via so-called correction sets. When all NMR spectra have been simulated, they are compared to the experimental spectrum and ranked after fit. The smallest deviation is normally obtained for the spectrum of the correct structure.

The results of various simulations of NMR spectra, which were performed in the menu 'Structural Determination', can be obtained in the menu 'Results' (and sublevels). These will be given as information on the proposed saccharide structures and the corresponding NMR spectra as lists of chemical shifts or as a bar graph NMR spectrum (Graphical Spectral Output). Via the menu 'CHEM-X Interface', an output file can be created by CASPER which can be read by the molecular modeling program CHEM-X.[11] The file contains a number of commands that can be interpreted by the molecular modeling program, which builds a 3D structure with reasonable geometry. The geometry of this 3D structure may then be further optimized.

Under the menu 'Simulation of Spectra of One Structure', the NMR chemical shifts for a specific oligo- or polysaccharide can be obtained.

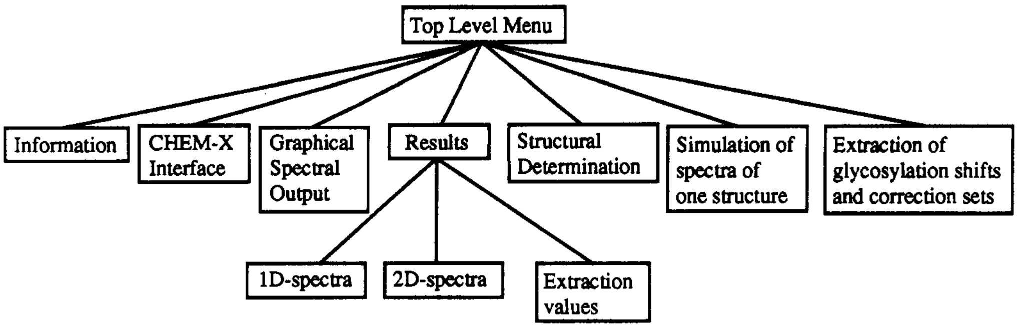
Glycosylation shifts and correction sets may be extracted from fully assigned NMR spectra found in the literature. This extraction can be performed under the menu 'Extraction of Glycosylation Shifts and Correction Sets'.

DATABASES

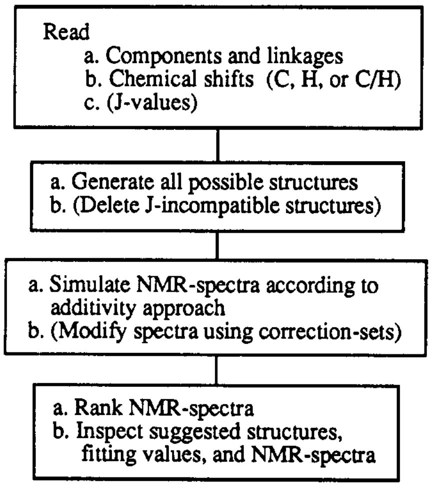
The CASPER program uses ^{13}C and/or ^{1}H NMR data. The data used in a simulation of an NMR spectrum can be divided into three categories: (i) the chemical shifts of the monosaccharides, (ii) the glycosylation shifts of disaccharides, i.e., the differences between the chemical shifts of the signals of

0095-2338/91/1631-0508$02.50/0

Scheme I. Overview of the Menus of CASPER


Top Level Menu
Information
CHEM-X Interface
Graphical Spectral Output
Results
Structural Determination
Simulation of spectra of one structure
Extraction of glycosylation shifts and correction sets
1D-spectra
2D-spectra
Extraction values

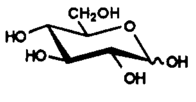
Scheme II. Generalized Structural Determination

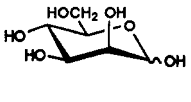

Read
a. Components and linkages
b. Chemical shifts (C, H, or C/H)
c. (J-values)
a. Generate all possible structures
b. (Delete J-incompatible structures)
a. Simulate NMR-spectra according to additivity approach
b. (Modify spectra using correction-sets)
a. Rank NMR-spectra
b. Inspect suggested structures, fitting values, and NMR-spectra

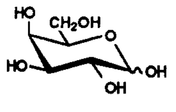
a disaccharide and of its constituent monosaccharides, and (iii) the correction sets which are also glycosylation shifts to be added in the simulation of NMR spectra of sterically strained structures. These correction values are the differences between the observed chemical shifts for sterically strained trisaccharide models and those calculated by the additivity approach.

In order to obtain maximum consistency, all spectra recorded by us were for solutions in D_2O at 70 °C, using 1,4-dioxane (δ_C 67.40) or sodium (trimethylsilyl)propionate-d_4 (TSP, δ_H 0.00) as the internal references. The elevated temperature is advantageous as it gives narrower lines in the polysaccharide NMR spectra.

Monosaccharides. CASPER contains at present data for 11 monosaccharides in their pyranosidic form. For each sugar, chemical shifts for both the α- and the β-anomeric form are present in the database. Furthermore, the methyl glycopyranosides of the same sugars are included. The monosaccharides are as follow: D-glucose, D-galactose, D-mannose, the corresponding uronic acids, the corresponding 2-acetamido-2-deoxy sugars, L-fucose (6-deoxy-L-galactose), and L-rhamnose (6-deoxy-L-mannose). In addition, data for D-glucose, D-galactose, and D-mannose with pyruvic acid ketalically linked to positions 4 and 6 are available, making a total of 14 monomers. The uronic acids and the pyruvylated sugars have sodium as their counter ion. The monosaccharide database contains both ^{13}C and ^{1}H NMR chemical shifts.


D-Glc


D-Man


D-Gal

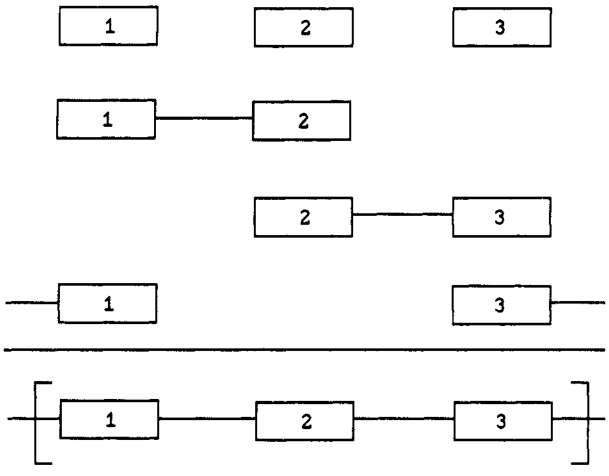
Glycosylation Shifts. The differences between the NMR chemical shifts of a disaccharide and those of the constituent monomers are termed the glycosylation shifts or Δ-set. Data for disaccharides with all possible combinations of the monosaccharides are included in the database except for those containing 4,6-pyruvylated sugar residues, which are treated as the corresponding monosaccharides with regard to the glycosylation shifts. An excerpt of the ^{13}C NMR glycosylation shift database with the glycosylation shifts is given below.

```
!ADGLC -4ADGLCN!
!880427, 3!
* 23 15
 7.66  0.25  0.11 -0.27  1.23 -0.25
-0.20 -0.21  0.25  7.76 -1.40 -0.08  0     0
 0.1
```

The glycosylation shifts may derive from experimental data on the actual disaccharide or be approximated from values of a similar disaccharide. The first line in the record contains the name of the disaccharide, ADGAL-4ADGLCN which is the abbreviated and latinized form of the disaccharide α-D-Gal*p*-(1→4)-α-D-Glc*p*NAc. The following line(s) include other comments, e.g., how an approximation is performed. The asterisk denotes the beginning of data to be read. The number following the asterisk is the index number of the disaccharide. The next number determines the number of values that should be read into the memory in a matrix. Two lines then include the glycosylation shifts of each constituent monosaccharide in the disaccharide (C1–C6, NAc—Me, C=O). The last figure is a check number (see below). The glycosylation shift database contains values for both ^{13}C and ^{1}H NMR spectra.

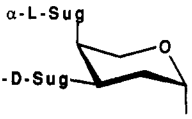
Correction Sets. When a branch point residue is vicinally disubstituted, i.e., 2,3- or 3,4-branched, changes in conformation and atomic interactions often occur with concomitant changes of chemical shifts. The chemical shifts will then no longer be the sum of the chemical shifts of the monosaccharides and the glycosylation shifts of the constituent disaccharides. The simple additivity approach is thus not enough. Therefore correction sets, i.e., another set of glycosylation shifts must be added. It has previously been shown[12,13] that for certain 2-linked sugar residues in a linear sequence and its neighboring residues deviations from the pure additivity approach may be observed depending on the vicinal disubstitution. The 2-linked sugar residue may thus also be treated as vicinally disubstituted though it is formally not. The number of combinations of correction sets possible from the 14 monosaccharides will be large. A reduction of the number of possibilities has been done by only allowing variations in the absolute and anomeric configuration of the substituting sugar residue, i.e., D/L and α/β, respectively. The branching sugar residue may either be an α or a β sugar having either the D or L configuration. Since only sugars with gluco, manno, and galacto configuration or derivatives thereof are represented in the database, the combination of substituted vicinal hydroxy groups in terms of axial or equatorial groups will be limited. Thus, the possible combinations are then equatorial–equatorial or axial–equatorial 2,3-linked and equatorial–equatorial or equatorial–axial 3,4-linked. As the sugars substituting the branching sugar residue may be α, or β and D or L, the number

Scheme III. Schematic Simulation of a Spectrum of a Polysaccharide



of combinations for branchpoint correction sets is 256. As example of a ^{1}H NMR branch point correction set is shown below.

```
!BD_AL_3,4AD__ea!
!880801!
* 227 16
-0.01 -0.03  0.02  0.01  0.00
 0.32 -0.05  0.00  0.01 -0.05
 0.03 -0.06  0.03  0.05  0.03
 0.01
```



The sugar linked to the lowest numbered carbon always has its values on the first line, and values for the branch point residue are on the third line. In order to simplify calculations, the same number of corrections is used for each ^{13}C and ^{1}H correction set. For ^{13}C NMR data the number of values for each monosaccharide is six, and for ^{1}H NMR data it is five.

When the sugar chains contain a 2-linked sugar residue, correction sets with values for sugar1, sugar2, and sugar3 are available for certain trisaccharide elements.

sugar1-(1→2)-sugar2-(1→X)-sugar3 X = 2, 3, 4

All these sugars may take the α, β, D, or L configuration. When sugar3 is 2- or 4-linked it is classified as either equatorially or axially substituted and when it is 3-linked it is always equatorially substituted. When sugar3 is 6-linked no correction is performed as it is assumed[14,15] that only small chemical shift changes are introduced from the distant 6-linked sugar residue. The number of correction sets is 640.

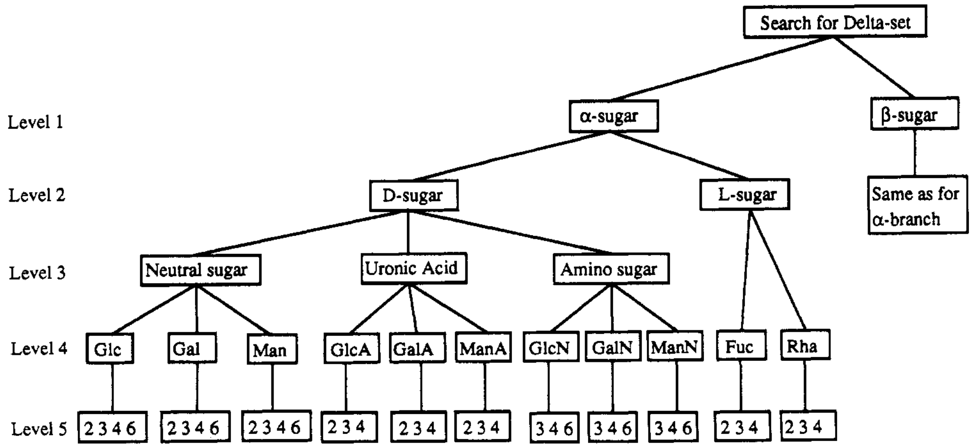
Check Numbers. A number of di- and tri-saccharides have been studied, and the signals in the NMR spectra assigned in order to obtain reliable glycosylation shifts. Glycosylation shifts for many disaccharides in the database will, however, be approximated from other values. The interresidue atomic interactions in the disaccharides α-D-Glc*p*-(1→4)-α-D-Glc*p* and α-D-Gal*p*-(1→4)-α-D-Glc*p* should be closely similar as the change of the 4′-hydroxyl group from equatorial to axial is far away from the glycosidic linkage. The chemical shifts of the two disaccharides will not be similar, but the glycosylation shifts will. Thus it is possible to use approximated values. A disaccharide with its NMR signals assigned is given a check number of 0.01. The above approximation is given a check number of 0.1. For more coarse approximations, check numbers of 1 or 10 are given. Similar check numbers are also present for the correction sets. They are termed steric interaction (SI) check numbers.

SIMULATION OF AN NMR SPECTRUM

Most glycosylation shifts appear at or next to the glycosidic linkage but in order to take advantage also of the small changes, all glycosylation shifts are used in the simulation. Furthermore, the database contains values for every possible disaccharide combination so that when data become available approximations can easily be exchanged.

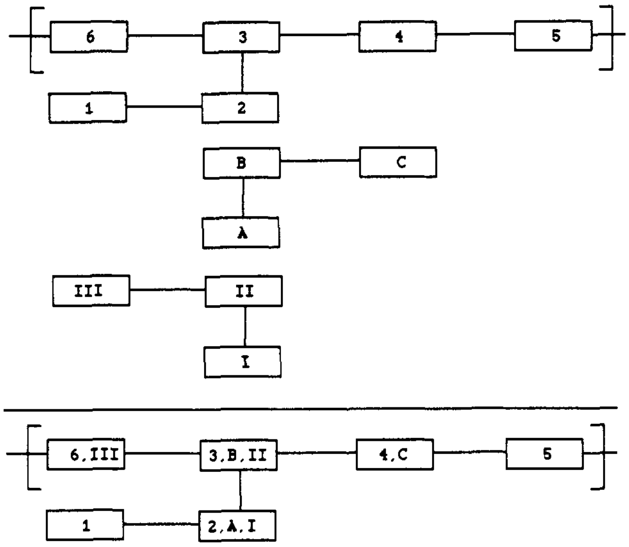
The chemical shifts for signals from an oligo- or polysaccharide are dependent on the constituent monosaccharides, their anomeric configuration, and the nature of the flanking residues. The simulation of an NMR spectrum of a polysaccharide is schematically depicted in Scheme III. First, information on the monosaccharides including their anomeric configuration in the repeating unit is supplied. Box number 1 in Scheme III then represents, e.g., α-D-glucopyranose. The ^{1}H or ^{13}C NMR chemical shifts for each monosaccharide are then read from the database and stored as a vector. Second, the glycosylation shifts for each glycosidic linkage are added. The identification of the first disaccharide element, 1→2, is done by a tree-search algorithm shown in Scheme IV. The first sugar residue is identified by a search from level 1 to level 4. Sugars with a 4,6-pyruvate substituent are treated as neutral sugars in this search as a substituent does not significantly interfere with the glycosidic linkage. The second sugar residue is identified in a repeated search. The linkage of the second sugar is identified at level 5, and the disaccharide element is thus identified, and the appropriate glycosylation shifts can be read and added to the chemical shifts of the two monosaccharides. The search for the second disaccharide element, 2→3, is made analogously, and the glycosylation shifts

Scheme IV. Search Algorithm for Location of a Δ-Set[a]



[a] For a disaccharide element, the first sugar is located at level 4. The procedure then identifies the second sugar (level 4) and its linkage (level 5), which is the linkage between the two sugars.

Scheme V. Simulated Spectrum Corrected for Interactions Due to Vicinal Disubstitution



are added to the chemical shifts of monosaccharides 2 and 3. As the polysaccharide is composed of repeating units, the last disaccharide element to be identified and added is 3→1, and after the addition of the appropriate glycosylation shifts, the NMR spectrum of the polysaccharide has been simulated. When an NMR spectrum of an oligosaccharide is simulated, the glycosylation shifts for the last disaccharide element, 3→1, are not added.

The simulation of the NMR spectrum of a branched polysaccharide is shown in Scheme V. If the branching point is substituted so that residues 2 and 6 are not vicinal, and no residue is 2-linked, the simple additive approach is used and six sets of glycosylation shifts are added to the chemical shifts of the monosaccharides. The only difference from the simulation in Scheme III is that the last residue (no. 6) is linked to the branch point residue (no. 3) instead of residue 1.

If the branching point is 2-linked a set of correction values is added, shown as boxes A, B, and C in Scheme V. These are added to residues 2, 3, and 4, respectively. In the example, the branched sugar residue number 3 is 2,3-linked, and the sugar residue number 4 is 2-, 3-, or 4-linked. The oligosaccharide belonging to the correction set for a 2-linked sugar residue is identified in a binary search (α/β, D/L) and added to monosaccharides 2, 3, and 4. If residue 3 is also 3-linked, i.e., 2,3-linked and thereby vicinally disubstituted, another set of correction values is added (I, II, and III), in this example to residues 2, 3, and 6, respectively. The branch point residue has thus been modified three times with glycosylation shifts and twice with correction sets.

An approach similar to that used for addition of correction sets could be used for, e.g., *O*-acetyl groups as some rules have been developed.[16] The same type of simulation could be performed for phosphate or sulfate substituents that are also commonly found as substituents in oligo- and polysaccharides.

Simulation of the C,H-correlation spectrum is performed by the creation of pairs of signals starting out from the simulated proton spectrum. For each proton signal, the appropriate carbon signal is identified, and these signals are stored pairwise.

STRUCTURAL DETERMINATION

Structural determination is the major aim of CASPER. A major object in the development of an automated approach for structural determination was simplicity. The input to CASPER consists of information on the constituent sugars and their linkage position(s) together with an experimental NMR spectrum.

From the top level menu, 'Structural Determination' can be chosen. The flow chart is shown in Scheme VI. The first time the simulation program is to be used, the databases are read from files. At all other times the start of a new simulation must be verified so that the old results should not be deleted accidentally. As spectral data for a structural determination either a one-dimensional or a two-dimensional NMR spectrum can be used. If a two-dimensional spectrum is used, a preliminary ranking of the ^{13}C NMR spectra is first performed in order to reduce the ranking list, as the comparison between experimental and simulated two-dimensional NMR spectra would be time-consuming to perform for all the possible structures. The experimental NMR spectrum may be read from a file or given interactively. Then a number of Boolean variables need to be set to define simulation parameters. The saccharide is defined as either an oligo- or a polysaccharide, and for the former it is necessary to define whether the oligosaccharide has a reducing end or is a methyl glycoside. It should then be stated whether the saccharide is linear or branched. Not more than one branching residue is allowed at present. The number of sugars in the repeating unit or in the oligosaccharide should then be entered (2–6 at present). Homopolysaccharides may be simulated by treating them as if they were polysaccharides with disaccharide repeating units. ^{13}C or ^{1}H NMR spectra or both may be simulated in one run. If the two-dimensional approach is to be chosen later, both the ^{13}C and ^{1}H NMR spectrum will be simulated.

The coupling constants of the anomeric proton signals can be used as constraints to reduce the number of possible combinations for a repeating unit. The $^{3}J_{H1,H2}$ values are divided into three categories: large (~8 Hz) for sugars with trans-diaxial H-1-H-2, e.g., as in β-D-glucopyranose; medium (~4 Hz) for sugars with an equatorial H-1 and an axial H-2, e.g., as in α-D-glucopyranose; or small (<2 Hz) for sugars with an equatorial/axial H-1 and an equatorial H-2, e.g., as in α/β-D-mannopyranose, respectively. The number of each type must then correlate in the simulated and experimental spectra. The one-bond C–H coupling constant for the signals from the anomeric carbon can be used in a similar way. A large value (~170 Hz) refers to α-glycosides and a small value (~160 Hz) to β-glycosides or, more generally, to glycosides with an equatorial and an axial anomeric hydrogen, respectively. Data on monosaccharide composition obtained from sugar analysis and the linkage position(s) identified from methylation analysis[17] are entered as input before the simulation begins.

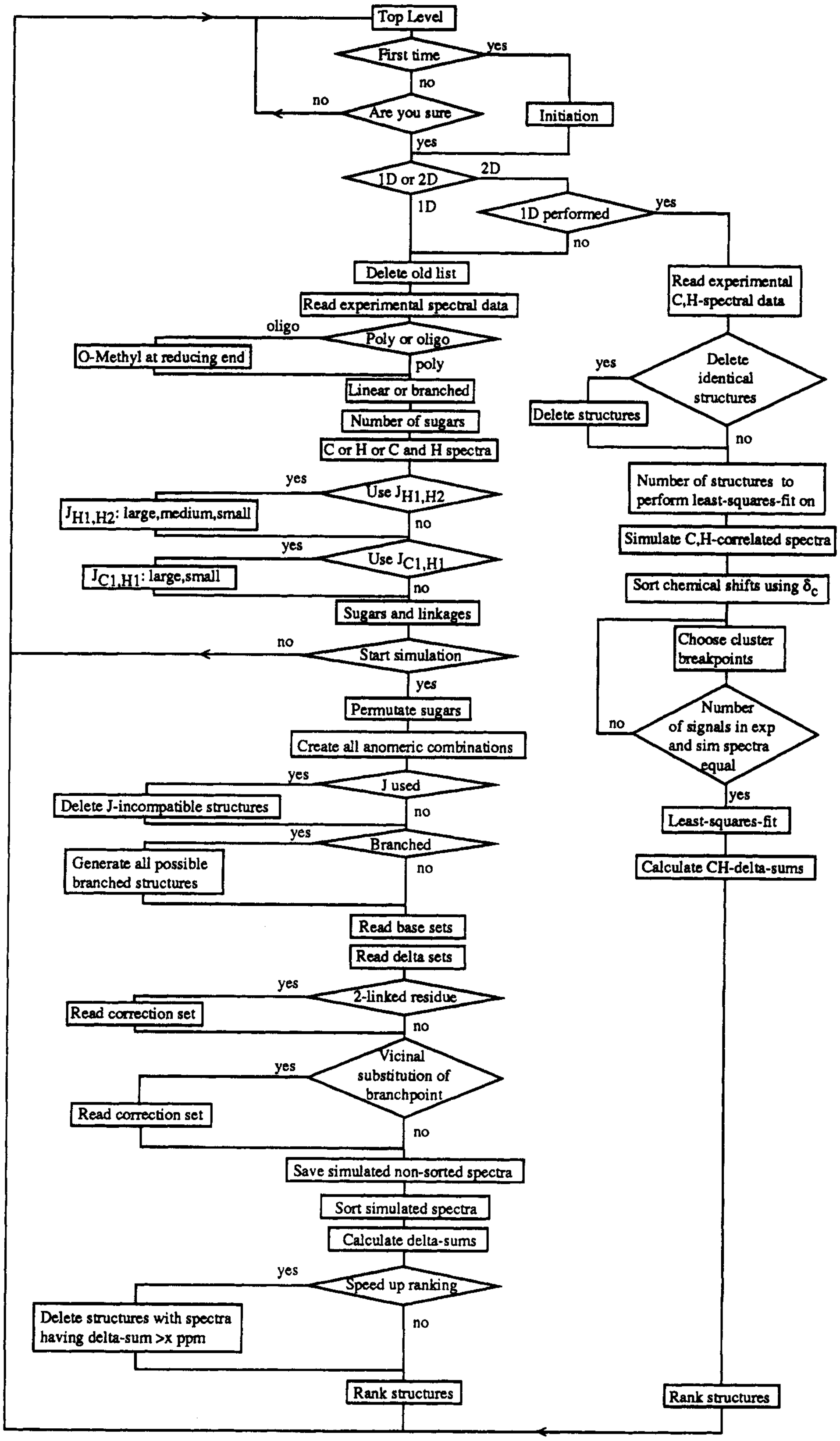
For the polysaccharide all permutations of sugars are created, which is $(n - 1)!$ for a repeating unit of n sugars. The oligosaccharide has $(n - 2)!$ possible permutations as the terminal, and the reducing sugar residues are defined by chemical means. The anomeric configuration may be either α or β so the number of possible combinations for any permutation is 2^n. The generation of structures also includes the generation of anomeric configurations. Therefore, the restraints from coupling constant information ($^{3}J_{H1,H2}{}^{1}J_{C1,H1}$) can be used to exclude structures, prior to spectral simulation, which are not consistent with the coupling information. The number of each type must then correlate in experimental and simulated spectra.

If a branch point residue is present, it may have the substituting residues arranged in two ways, e.g., 3 and 4 or 4 and 3. This doubles the number of possible structures. The number of possible structures m of a polysaccharide repeat with n sugars is then

$$m = (n-1)!\times 2^{(n+b)}; \qquad b = \begin{cases} 0 \text{ if linear} \\ 1 \text{ if branched} \end{cases}$$

For each structure, an NMR spectrum will be simulated as described in Schemes III and V. The simulated spectra are

Scheme VI. Flowchart of Structural Determination



now stored with the chemical shifts of each sugar residue saved separately so that they can be identified. The simulated spectra are sorted in decreasing order of chemical shift, i.e., as in the experimental spectrum and compared on a signal-to-signal basis to the experimental spectrum. The absolute values of each difference between the experimental and simulated signals are added to a so-called Δ-sum. A low Δ-sum indicates a good fit between the experimental spectrum and a simulated spectrum, whereas a high Δ-sum indicates a bad fit. When a low average Δ-sum is obtained, i.e., <0.2 ppm/signal for ^{13}C and <0.02 ppm/signal for ^{1}H, suggested structures with a Δ-sum twice the lowest one can be regarded as highly unlikely and need not be considered in further analysis. When a structural determination of a one-dimensional spectrum is performed in this way, the results can be obtained under the 'Results' menu.

The analysis using two-dimensional NMR data takes advantage of the ranked suggestions from the one-dimensional

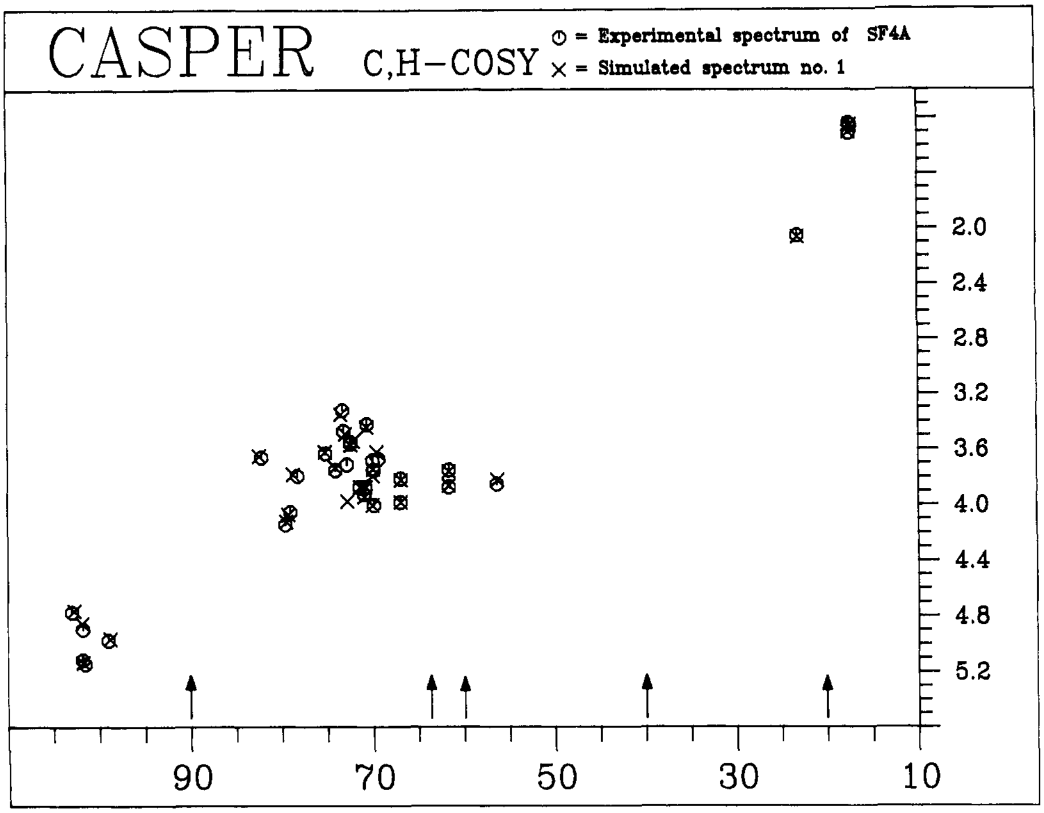


Figure 1. Graphical output of the carbon–proton correlation spectrum of the O-antigen polysaccharide from *Shigella flexneri* type 4a and the simulated spectrum with the best fit. The arrows below the picture show different regions into which the spectra are divided in order to lessen the analysis time of the least-squares-fit procedure.

run using ^{13}C NMR data. The experimental C,H-correlation spectrum is read in which the C,H pairs are stored in decreasing ^{13}C NMR chemical shift order. Structures with identical residues are only analyzed once, which also reduces simulation time. The experimental and simulated two-dimensional NMR spectra will be compared by using a least-squares-fit procedure (see below) but as this procedure is time-consuming for large data sets, it is preferably performed on a limited number of structures (~5–10 structures). The C,H pairs of the simulated spectra are sorted in decreasing order of the ^{13}C NMR chemical shifts so that a faster least-squares fit can be performed.

To speed up calculations, the two-dimensional C,H-correlation spectrum is divided into regions or clusters by so-called cluster breakpoints as shown in Figure 1. A region contains signals from similar types of carbons and protons, and the different regions are indicated by arrows in Figure 1. Signals from the anomeric carbons are found at $\delta > 90$, those for ring carbons at δ 90–63, those for unsubstituted hydroxymethyl carbons at δ 63–60, those for carbons bearing a nitrogen group at δ 60–40, the *N*-acetyl methyl group at δ 40–20, and the methyl groups of 6-deoxy sugars at δ 20–0. Furthermore it is possible, and advisable, to divide the region for ring carbon signals into at least two regions as for a hexasaccharide there will be at least 36 C,H pairs which need to be least-square-fitted. This division of the ring carbon region should preferably be made where the ^{13}C NMR spectrum contains gaps. In the experimental and simulated NMR spectrum, each region must have the same number of C,H pairs. If this is not the case, new cluster breakpoints have to be chosen.

For each region a least-squares fit between experimental and simulated signals is performed. The best fit for a region has been found when

$$\sum_{i}^{n}(x_i^2 + y_i^2) > \sum_{i}^{n_{\max}}(x_i^2 + y_i^2)$$

where $n_{\max}$ equals the number of signals in a cluster; n has the range $1 < n \leq n_{\max}$; x equals the ^{13}C NMR chemical shift difference of a simulated signal and an experimental signal; and y equals the ^{1}H NMR chemical shift difference of the respective signals multiplied by 10 in order to obtain the same order of magnitude for data in both dimensions. The C,H Δ-sum is calculated for the best fit of each simulated structure by

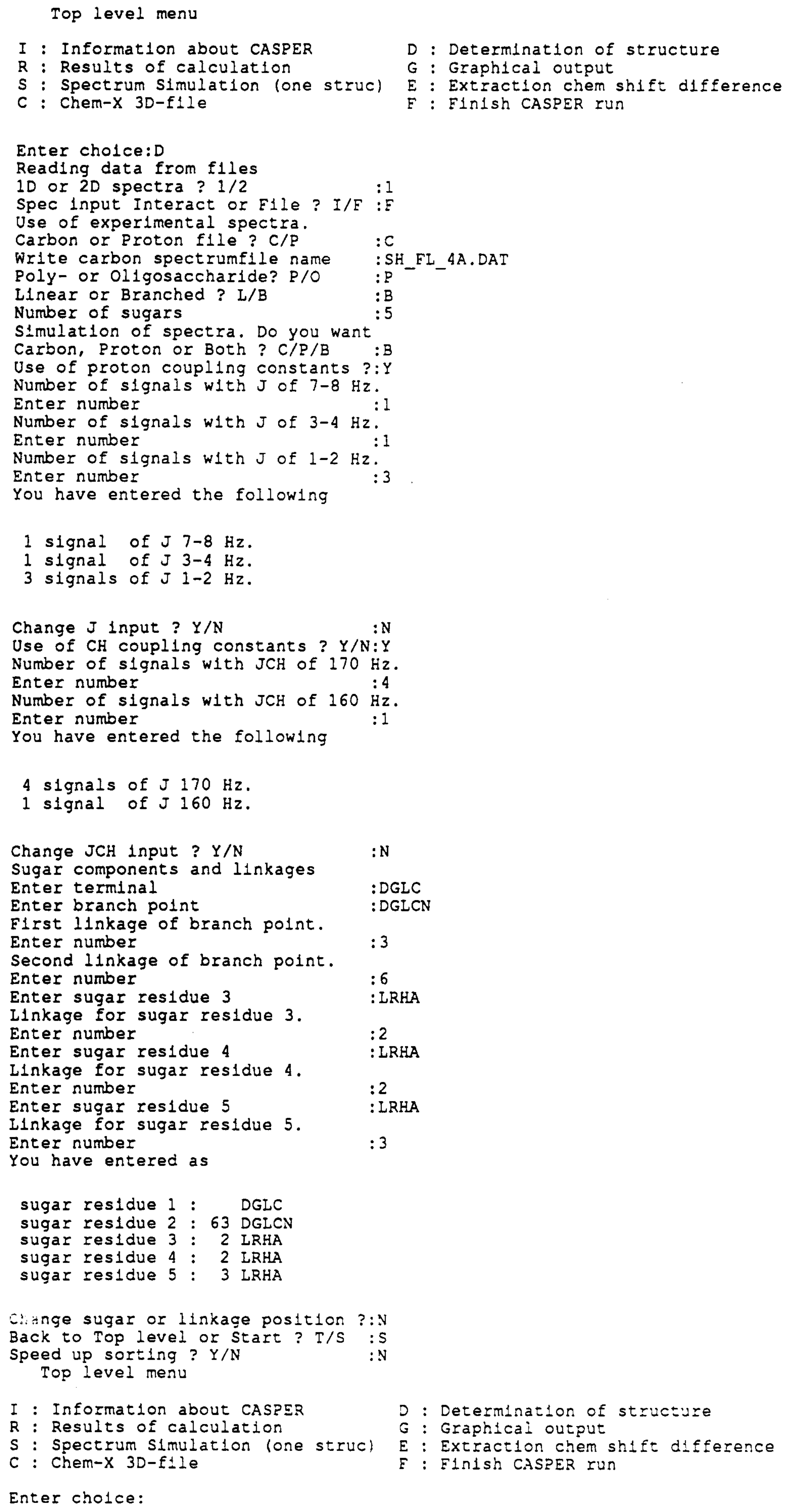
$$[\sum_{i}^{\max}(x_i^2 + y_i^2)]^{1/2}$$

where max is the number of C,H pairs in the spectrum and x and y are the same as above. The different structures are then ranked after increasing C,H Δ-sums and a structural determination using a two-dimensional spectrum have been performed. The results can be obtained under the 'Results' menu.

EXTRACTION OF GLYCOSYLATION SHIFTS AND CORRECTION SETS

The glycosylation shifts and correction sets of the database mostly come from studies on synthetic di- and trisaccharides. Structural studies of poly- and oligosaccharides often contain NMR spectra with partial or complete assignments although measurement conditions are not always equal to those most suitable for CASPER. These NMR data can be used by CASPER in order to obtain new sets of glycosylation shifts or correction sets. In principle, the procedure is the reverse of a structural determination.[18] A structure and its assigned spectrum are entered. A spectrum of the structure is simulated but the information sought for is omitted, i.e., a Δ-set or a correction set. The values are obtained by subtraction of the partially simulated spectrum from the assigned experimental spectrum. This procedure is especially useful for obtaining correction sets. The extracted data are at present entered manually into the

Scheme VII



```
    Top level menu

 I : Information about CASPER         D : Determination of structure
 R : Results of calculation           G : Graphical output
 S : Spectrum Simulation (one struc)  E : Extraction chem shift difference
 C : Chem-X 3D-file                   F : Finish CASPER run


 Enter choice:D
 Reading data from files
 1D or 2D spectra ? 1/2             :1
 Spec input Interact or File ? I/F :F
 Use of experimental spectra.
 Carbon or Proton file ? C/P        :C
 Write carbon spectrumfile name     :SH_FL_4A.DAT
 Poly- or Oligosaccharide? P/O      :P
 Linear or Branched ? L/B           :B
 Number of sugars                   :5
 Simulation of spectra. Do you want
 Carbon, Proton or Both ? C/P/B     :B
 Use of proton coupling constants ?:Y
 Number of signals with J of 7-8 Hz.
 Enter number                       :1
 Number of signals with J of 3-4 Hz.
 Enter number                       :1
 Number of signals with J of 1-2 Hz.
 Enter number                       :3
 You have entered the following


  1 signal  of J 7-8 Hz.
  1 signal  of J 3-4 Hz.
  3 signals of J 1-2 Hz.


 Change J input ? Y/N               :N
 Use of CH coupling constants ? Y/N:Y
 Number of signals with JCH of 170 Hz.
 Enter number                       :4
 Number of signals with JCH of 160 Hz.
 Enter number                       :1
 You have entered the following


  4 signals of J 170 Hz.
  1 signal  of J 160 Hz.


 Change JCH input ? Y/N             :N
 Sugar components and linkages
 Enter terminal                     :DGLC
 Enter branch point                 :DGLCN
 First linkage of branch point.
 Enter number                       :3
 Second linkage of branch point.
 Enter number                       :6
 Enter sugar residue 3              :LRHA
 Linkage for sugar residue 3.
 Enter number                       :2
 Enter sugar residue 4              :LRHA
 Linkage for sugar residue 4.
 Enter number                       :2
 Enter sugar residue 5              :LRHA
 Linkage for sugar residue 5.
 Enter number                       :3
 You have entered as


  sugar residue 1 :    DGLC
  sugar residue 2 : 63 DGLCN
  sugar residue 3 :  2 LRHA
  sugar residue 4 :  2 LRHA
  sugar residue 5 :  3 LRHA


 Change sugar or linkage position ?:N
 Back to Top level or Start ? T/S  :S
 Speed up sorting ? Y/N             :N
    Top level menu

 I : Information about CASPER         D : Determination of structure
 R : Results of calculation           G : Graphical output
 S : Spectrum Simulation (one struc)  E : Extraction chem shift difference
 C : Chem-X 3D-file                   F : Finish CASPER run

 Enter choice:
```

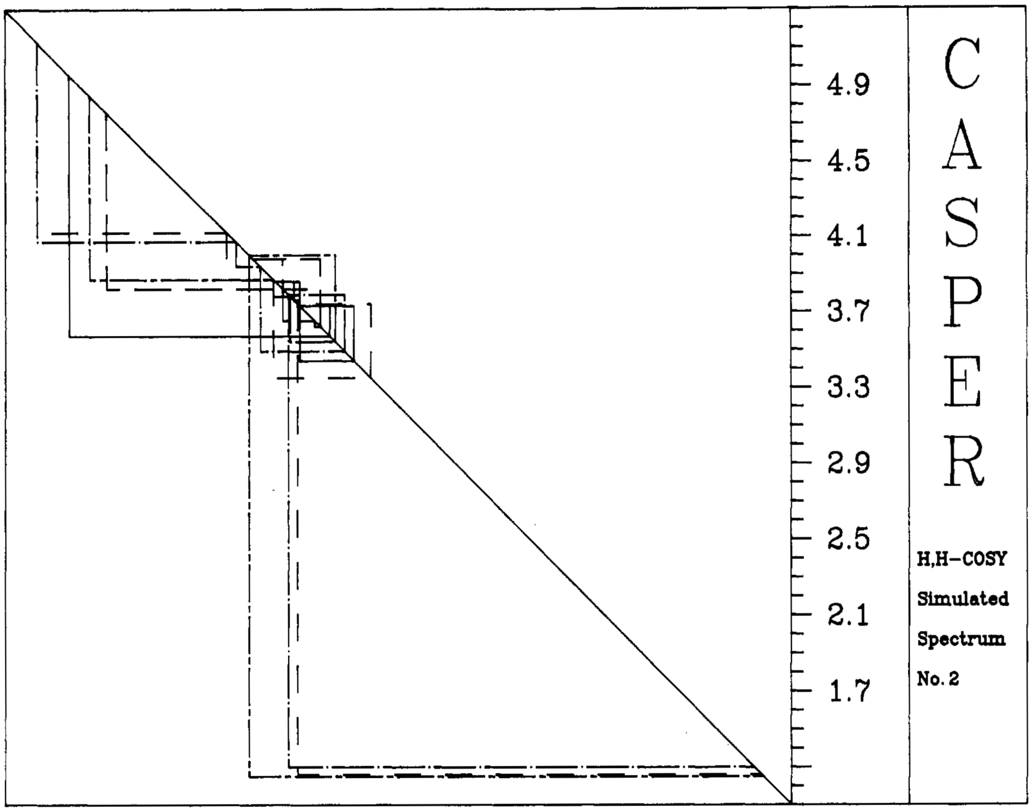


Figure 2. H,H-COSY spectrum of simulated spectrum number 2. The spin–spin connectivity pathway of each spin system is shown by different lines.

database but an automatic procedure is planned. This procedure will average old and new data and therefore improve data.

EXAMPLE AND OUTPUT

An example of a structural determination is given here. The structure of the O-polysaccharide from *Shigella flexneri* type 4a is shown below.

```
→3)-β-D-GlcpNAc-(1→2)-α-L-Rhap-(1→2)-α-L-Rhap-(1→3)-α-L-Rhap-(1→
        6
        ↑
        1
  α-D-Glcp
```

This structure has been determined by chemical[19] and verified by spectroscopic[14,15] methods. The data given to CASPER are entered as shown in Scheme VII. This example shows the use of both one- and two-dimensional data. The one-dimensional analysis always comes first, the input of which is shown in the scheme. After definition of the type of saccharide analyzed, i.e., if it is an oligo- or polysaccharide and if it is linear or branched, the user is asked whether carbon and/or proton data are required and if coupling constants are to be given. When information on sugars and linkages is given, the analysis will proceed as described above. The output of the three structures with lowest Δ-sum is given in Scheme VIII in which duplicate structures arising from the presence of identical residues have been omitted. The spectral data can be given in a numerical form in decreasing order for both ^{13}C and ^{1}H NMR chemical shifts. The ^{13}C spectra may also be shown in graphical output mode as a bar graph ^{13}C NMR spectrum. Two ^{13}C NMR spectra can also be compared to each other in a split-screen mode. The graphical output can be in color or monochrome. When two signals overlap in the color mode, the color of the signals is changed. Spectra sorted by residue (C1–C6, Me, C=O) can also be obtained corresponding to the calculated chemical shifts, i.e., tentatively assigned spectra for a structure (Scheme VIII). These spectra can also be used to simulate the H,H-COSY spectrum. The spectrum of structure 2 is shown in Figure 2 in which the spin–spin connectivity pathway of each spin system is indicated by different lines. Only half the number of the cross-peaks is shown in order to reduce spectral overlap.

Scheme VIII. Output of a Structural Determination Based on ^{13}C NMR Data of the O-Antigen Polysaccharide from *Shigella flexneri* Type 4a (SF4a) Including Ranked Structures and Their Δ-Sums and Check Numbers and Numerical Output of Nonsorted Simulated Spectrum of Simulated Structure Number 2

```
      SH_FL_4A
No.   Polysaccharide.

 2  -3BDGLCN-2ALRHA -2ALRHA -3ALRHA -
      6
    ADGLC

 4  -3BDGLCN-2ALRHA -3ALRHA -2ALRHA -
      6
    ADGLC

 6            -6BDGLCN-2ALRHA -3ALRHA -
                 3
    ADGLC -2ALRHA


No.    13C      13C Sum      13C      13C-SI
     Deltasum    /sig      Check#    Check#

 2     3.5      0.11       0.14      20.00
 4     5.3      0.16       0.14      20.00
 6     7.8      0.24       0.23      20.00

J12 = 113 and JCH = 41 used to eliminate structures


Spectrum number      2.
  98.8   72.4   74.1   70.7   72.8   61.7
 102.8   56.3   82.5   69.6   75.2   66.9   23.1 175.0
 101.8   79.5   70.8   73.5   70.0   17.5
 101.8   79.3   70.9   73.0   70.0   17.6
 102.0   71.4   78.8   72.2   70.0   17.4
```

The structural determination may continue from the ranked structures obtained from the simulation of one-dimensional spectra. The three best-fitting structures are chosen for

Scheme IX. Output of a Structural Determination Based on C,H-Correlation Data of SF4a

```
     SF4A
  No. Polysaccharide.

   1 -3BDGLCN-2ALRHA -2ALRHA -3ALRHA -
         6
       ADGLC

   2 -3BDGLCN-2ALRHA -3ALRHA -2ALRHA -
         6
       ADGLC

   3          -6BDGLCN-2ALRHA -3ALRHA -
                  3
       ADGLC -2ALRHA


  No.   C,H       13C      1H     13C-SI      1H-SI
     Deltasum   Check#   Check#   Check#     Check#

   1    1.5     0.14     0.14    20.00       20.00
   2    4.3     0.14     0.14    20.00       20.00
   3    5.4     0.23     1.13    20.00       20.00

 J12 = 113 and JCH = 41 used to eliminate structures

 Experimental C,H-correlation spectrum.
  103.0 4.77 101.8 5.11 101.8 4.89 101.6 5.14  99.1 4.97
   82.2 3.66  79.6 4.14  79.1 4.05  78.3 3.79  75.2 3.63
   74.1 3.75  73.3 3.32  73.2 3.47  72.8 3.71  72.4 3.56
   72.4 3.55  71.5 3.87  70.9 3.92  70.8 3.87  70.7 3.42
   70.0 3.68  69.9 3.75  69.9 4.00  69.4 3.67  67.0 3.81
   67.0 3.98  61.7 3.75  61.7 3.87  56.4 3.85  23.2 2.05
   17.6 1.31  17.5 1.24  17.4 1.25

 Spectrum number      1.
  102.8 4.76 101.8 5.13 102.0 4.85 101.8 5.13  98.8 4.96
   82.5 3.65  79.5 4.12  79.3 4.07  78.8 3.78  75.2 3.62
   74.1 3.73  73.5 3.35  73.0 3.49  72.8 3.73  72.4 3.57
   72.2 3.54  71.4 3.87  70.9 3.94  70.8 3.87  70.7 3.44
   70.0 3.74  70.0 3.79  70.0 4.00  69.6 3.62  66.9 3.82
   66.9 3.98  61.7 3.76  61.7 3.86  56.3 3.82  23.1 2.06
   17.6 1.30  17.5 1.26  17.4 1.25

 Spectrum number      2.
  102.8 4.76 102.8 4.96 100.9 4.96 101.9 5.19  98.8 4.96
   82.5 3.65  79.5 4.12  78.7 3.86  79.8 3.80  75.2 3.62
   74.1 3.73  73.3 3.35  72.9 3.46  72.8 3.73  72.4 3.57
   72.5 3.57  70.9 3.83  70.8 4.14  70.8 3.90  70.7 3.44
   69.9 3.77  70.1 3.81  70.0 3.95  69.6 3.62  66.9 3.82
   66.9 3.98  61.7 3.76  61.7 3.86  56.3 3.82  23.1 2.06
   17.4 1.29  17.5 1.25  17.5 1.27
```

analysis by using C,H-correlation spectra. This is performed as described above. The output is shown in Scheme IX in which the structures have been ranked on their C,H Δ-sum. The numerical form of the experimental spectrum and the two spectra with best fit are also shown. The graphical output of the experimental C,H-correlated spectrum and that with best fit is shown in Figure 1.

A structural determination may also be based on a ^{1}H NMR spectrum instead of a ^{13}C NMR spectrum. The chemical shifts are preferably obtained from an H,H-COSY spectrum and entered in decreasing order in the input spectrum. Not all ^{1}H NMR signals need to be obtained in order to perform an analysis, but the same number of signals from each sugar residue, e.g., from H1, H2 and H3, should be used. All numerical output may also be written directly to a file.

CONCLUSIONS

The reliability of CASPER has been established, and reports of structural determinations based on a computer approach have started to appear.[20,21] The quality of the database is steadily increasing and with the advent of "inverse-detection" in NMR spectroscopy, the analysis of small amounts of carbohydrate material will be possible.[22] In the future CASPER will address more complex carbohydrate structures, e.g., oligosaccharides from glycoproteins. The development of new algorithms based on artificial intelligence and the interlinking of CASPER with various other techniques such as mass spectrometry could develop it into a system for total analysis of carbohydrates.

TECHNICAL DESCRIPTION OF THE PROGRAM SYSTEM

CASPER is written in Vax-11 Pascal and runs on a VAX 11/750 computer. The menu-driven program consists of some 150 procedures in six modules which are linked together into an executable program. The size of the Pascal program is approximately 15 000 lines. The data type used for CASPER is a linked list where each cell represents a structure. The information to be stored in a cell is implemented as a *record-type* which facilitates easy addition of further fields such as those for coupling constants or Δ-sums.

The four monosaccharide databases each have 28 entries of 5–10 numbers. The two databases of glycosylation shifts each have 1584 entries of 11–17 numbers. Two of the correction-set databases have 256 entries and two have 640 entries of 16–19 numbers. The size of the complete database is approximately 30 000 lines.

The graphical procedures are all programmed in Pascal, which facilitates easy transportation to various graphical packages, and then linkage with the GPGS-F package (Norwegian Association for Computer Graphics at the Computing Centre, University of Trondheim). Graphic output devices supported by version 2.1 of CASPER are Tektronics 4105 or higher versions or the Hewlett-Packard-7550 pen-plotter for paper drawings.

The output file for molecular modeling produced by CASPER is directly readable by the CHEM-X program.

ACKNOWLEDGMENT

This work was supported by grants from the Swedish Natural Science Research Council and the Swedish National Board for Technical Development.

REFERENCES AND NOTES

(1) Gray, N. A. B. *Prog. Nucl. Magn. Reson. Spectrosc.* **1982**, *15*, 201–248.

(2) Egli, H.; Smith, D. H.; Djerassi, C. *Helv. Chim. Acta* **1982**, *65*, 1898–1919.

(3) Hounsell, E. F.; Wright, D. J. *Carbohydr. Res.* **1990**, *205*, 19–29.

(4) Bot, D. S. M.; Cleij, P.; van't Klooster, H. A.; van Halbeek, H.; Veldink, G. A.; Vliegenthart, J. F. G. *J. Chemom.* **1988**, *2*, 11–27.

(5) Anderson, D. R.; Grimes, W. J. *Anal. Biochem.* **1985**, *146*, 13–22.

(6) Lipkind, G. M.; Shashkov, A. S.; Knirel, Y. A.; Vinogradov, E. V.; Kochetkov, N. K. *Carbohydr. Res.* **1988**, *175*, 59–75.

(7) Cumming, D. A.; Hellerqvist, C. G.; Touster, O. *Carbohydr. Res.* **1988**, *179*, 369–380.

(8) Jansson, P.-E.; Kenne, L.; Widmalm, G. *Pure Appl. Chem.* **1989**, *61*, 1181–1192.

(9) Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1987**, *168*, 67–77.

(10) Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1989**, *188*, 161–191.

(11) CHEM-X, developed and distributed by Chemical Design Ltd., Oxford, England.

(12) Adeyeye, A.; Jansson, P.-E.; Kenne, L.; Widmalm, G. Submitted to *J. Chem. Soc., Perkin Trans. 2* **1991**, 963–973.

(13) Hermansson, K.; Jansson, P.-E.; Kenne, L.; Widmalm, G.; Lind, F. Manuscript in preparation.

(14) Jansson, P.-E.; Kenne, L.; Wehler, T. *Carbohydr. Res.* **1987**, *166*, 271–282.

(15) Jansson, P.-E.; Kenne, L.; Wehler, T. *Carbohydr. Res.* **1988**, *179*, 359–368.

(16) Jansson, P.-E.; Kenne, L.; Schweda, E. *J. Chem. Soc., Perkin Trans. 1* **1987**, 377–383.

(17) Jansson, P. E.; Kenne, L.; Lindberg, B.; Liedgren, H.; Lönngren, J. *Chem. Commun., Univ. Stockholm* **1976**, *No 8*, 1–75.

(18) Jansson, P.-E.; Kenne, L.; Widmalm, G. *Acta Chem. Scand.* **1991**, *45*, 517–522.

(19) Kenne, L.; Lindberg, B.; Petersson, K.; Katzenellenbogen, E.; Romanowska, E. *Eur. J. Biochem.* **1978**, *91*, 279–284.

(20) Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1989**, *193*, 322–325.

(21) Baumann, H.; Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1991**, *211*, 183–190.

(22) Byrd, R. A.; Egan, W.; Summers, M. F.; Bax, A. *Carbohydr. Res.* **1987**, *166*, 47–58.