

Design Studies Related to the Development of Distributed Web-based European Carbohydrate Databases

*Design Study DS4: Rapid computer assisted
interpretation of NMR-spectra.*

Task Title:

DS4-T4: Development of algorithms for automatic NMR-spectra interpretation.

Deliverable:

DS4-D3: Report: "Algorithms for automatic NMR-spectra interpretation".

Due date of deliverable: 31.03.2008; Actual submission date: 18.04.2008

Start date of project: 01.04.2005; Duration: 48 months

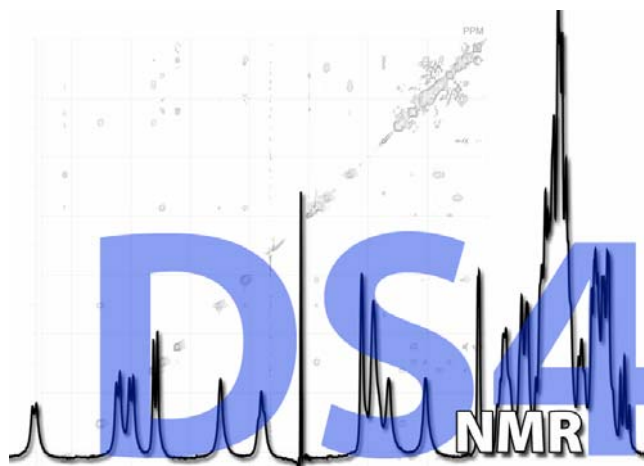
Organisation name of lead contractor for this deliverable:

Stockholm University, Stockholm,
Sweden

Authors

Dr. Göran Widmalm, Magnus Lundborg,
Dr. Bas Leeflang

Partners involved: 2, 3, 7



1 Algorithms for automatic NMR-spectra interpretation

1.1 NMR spectroscopy of carbohydrates

NMR spectroscopy is one of the most versatile techniques for investigation of molecular structure. The liquid state NMR spectroscopy technique is able to deliver atomic resolution information on carbohydrate molecules ranging from small and flexible monosaccharides to large polysaccharides, but also on complexes between carbohydrate molecules and other molecular systems. In addition, and highly advantageous, it is a non-destructive technique. However, carbohydrates are notorious for the limited resolution of their NMR spectra, due to similar chemical environments of the constituent atoms. Assignments of NMR resonances to the corresponding nuclei are therefore very tedious as well as prone to erroneous interpretations. As a result procedures and algorithms for automatic NMR-spectra interpretation are needed.

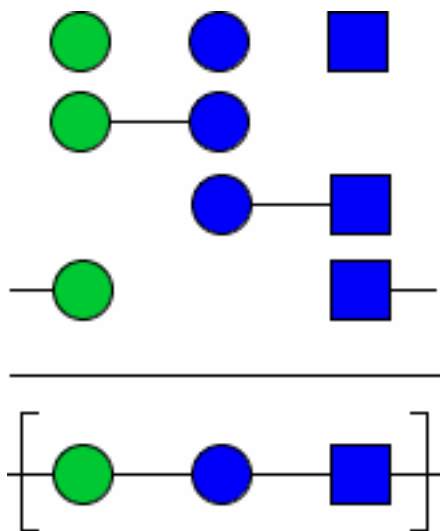
The methods presently used for determining the structure of complex carbohydrates depends largely on the level of detail desired and the amount of material available. Mass spectrometry and/or HPLC combined with exoglycosidase digestions are frequently used in the study of glycoproteins. The amount of material is often limited, but the number of possible structures can be restricted by biosynthetic considerations. In the analysis of polysaccharides of bacterial origin the amount of material available is quite often considerably larger, thereby facilitating the use of NMR spectroscopy as the method of choice. In elucidation of the carbohydrate structure, data-base approaches or computer programs for structure generation/spectrum calculation may be employed.

1.2 CASPER: an approach based on increment rules

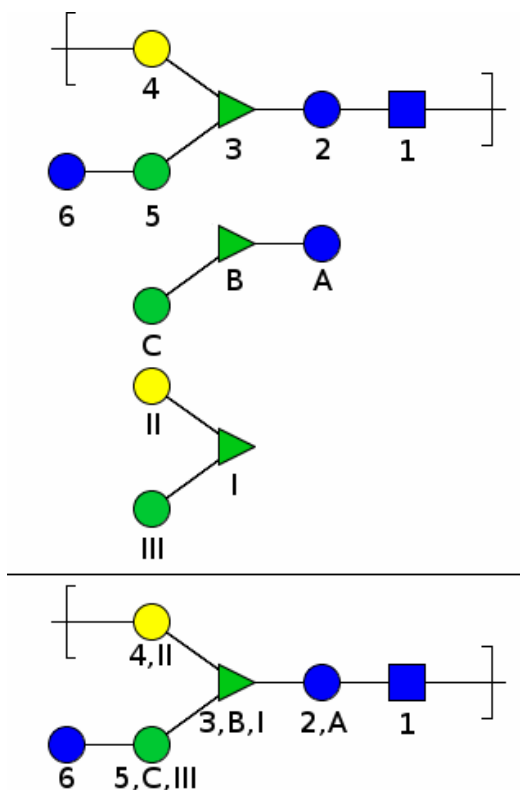
The computer program CASPER, Computer Assisted SPectrum Evaluation of Regular polysaccharides, belongs to the second group. The calculation of NMR-spectra can then be used for structure determination if it is combined with a suitable mechanism for generating trial structures. All structures consistent with data from chemical analyses (if present) are generated, their NMR-spectra calculated and finally ranked according to their agreement with the experimental spectrum. The computation of the chemical shifts is based on schemes with substituent-induced chemical shifts, referred to as glycosylation shifts. Additional NMR experiments may be required to distinguish between the suggested structures. The analysis is aided by the tentative assignments that are a result of this procedure. The application can make use of ^1H , ^{13}C and HSQC spectra. If only one spectrum is available ^{13}C NMR chemical shifts are preferred, since compared to ^1H the dispersion of the resonances is larger, the resolution is generally better and the predictability of the chemical shifts is higher.

The NMR spectra of oligo- and polysaccharides can be approximated from the chemical shifts of the constituent monosaccharides and the glycosylation shifts that are caused by substitution. The size of

the glycosylation shifts depends on the stereochemistry at and near the glycosidic bond, which allows the same glycosylation shifts to be used for the linkage between several different monosaccharides that have similar stereochemistry at the glycosidic linkage. Provided that only short range interactions are present, additivity of glycosylation shifts is observed, schematically shown below.



Exceptions are observed for residues with vicinal substitution, for example for branch points as well as for the substituting sugars. The vicinal substitution causes steric interactions not present in the corresponding disaccharide elements. To some extent this can be compensated for by the inclusion of corrections based on data from similar trisaccharides, schematically shown below.



The NMR data recorded to create the database use standardized conditions as far as possible.

A comparison between the simulated and experimental spectrum can be performed in different ways. The method employed in CASPER is a simple line-by-line comparison of the sorted spectra. The sum of the absolute values of the differences is then used for ranking. If the spectra differ in the number of resonances a recursive loop is used, which means that the procedure gets slower the higher the fraction of missing shifts. Root-mean-squared errors are also calculated for comparison with other programs. Matching simulated 2D spectra with the experimental spectrum is more complicated since the coordinates cannot be unambiguously sorted. This means that recursive calls are used when there can be different matching combinations. Since 2D one-bond $^{13}\text{C},^1\text{H}$ correlated NMR spectra can give information about connectivity, such information is used when assigning 1D spectra too. This means that 1D spectral assignments might be different when there is 2D correlation information included, in order to keep coherency with the procedure for atom assignments.

The anomeric $^3J_{\text{H,H}}$ or $^1J_{\text{C,H}}$ values can be used to reduce the number of generated structures. Only the unambiguously classified coupling constants are used as lower limits for the number of residues of a given type. This allows the use of $^3J_{\text{H,H}}$ or $^1J_{\text{C,H}}$ values even when some are unresolved in the spectrum or fall between two ranges. Structures violating these restrictions are not generated. In the case of structures containing many residues the use of these constraints may significantly reduce the time required for the calculations. However, the chemical shifts are normally sufficient to determine the anomeric configurations.

CASPER returns a list of the ten structures with the lowest calculated deviation from the experimental spectrum. From this list structures can be selected for closer examination. CASPER analysis results in a highly probable structure and as such it forms the basis of a rapid approach to final structure using a limited amount of additional information or just a candidate that can be used in biochemical, genetics or bioinformatics investigations.

1.3 ProSpectND: implementation of a flexible NMR spectral simulation module as a tool for fine-tuning

The computer program ProSpectND is a versatile NMR spectra processing and visualisation package that is available on all major ICT platforms, notably MS-Windows, Mac-OSX (PPC and Intel) and Unix based platform as Linux. This software is mature and is under continues development in the frame of EUROCarbDB. One aspect of these developments that is relevant to the NMR spectral assignment process is the facility to simulate 1D NMR spectra on the basis of rough assignment data (Chemical shift and J-couplings) and information on the characteristics of the used NMR spectrometer and the acquisition conditions of the experimental data.

NMR assignments are normally obtained through the acquisition and analysis of 1D and 2D NMR spectra. CASPER provides excellent opportunities to contribute in the assignment process. However,

when chemical shifts are very similar and thus when shifts are overlapping, CASPER is not discriminative enough to help the spectroscopist to make an assignment decision.

When chemical shifts are almost identical and the spins involved are also coupled (J), the multiplet structures of this spin system can be strongly deviant from the first-order shape that a spectroscopist is most comfortable with. This is often referred to as 'strong-coupling'. In these circumstances the spectral simulation can play a powerful role.

In addition to the simulation facility, ProSpectND provides an automated optimisation of the spectral parameters involved using an automated iterative procedure. Using the manual and automated optimisation procedures, ProSpectND provides the tools for the spectroscopist in the fine-tuning of the assignments in cases where the solution is not obvious with a simple visual inspection of the spectrum.