

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-T3: Agreements on quality measures for NMR data with enter the database.

Deliverable:

DS4-D2: Report: "Quality measures for NMR-spectra".

Due date of deliverable: 31.03.2007; Actual submission date: 31.03.2007

Start date of project: 01.04.2005; Duration: 48 months

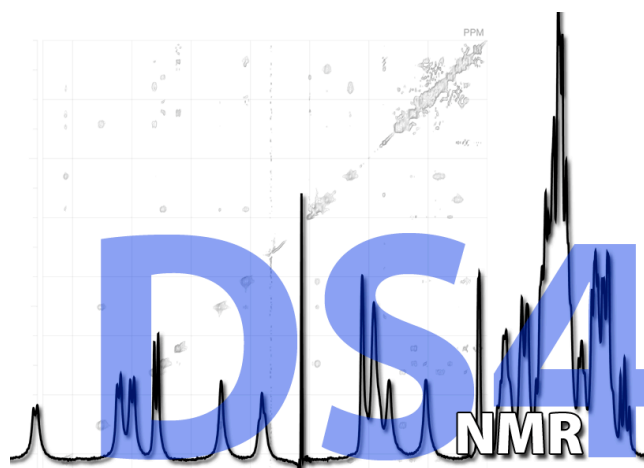
Organisation name of lead contractor for this deliverable:

Stockholm University, Stockholm,
Sweden

Authors

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

Partners involved: 2, 3, 7



1 NMR

1.1 Introduction

The goal of EUROCarbDB is to establish a framework to analyse, archive and retrieve experimental data on carbohydrates. NMR spectroscopy has proven to be a suitable tool to determine the primary sequence of glycans in combination with data on the monosaccharide composition and branching patterns. NMR can be used as well for more in-depth 3D structure and dynamics determinations.

This report is the outcome of an evaluation on what quality marks and criteria EUROCarbDB should be used for NMR data on carbohydrates. On the one hand there is a clear desire to have well defined quality criteria for NMR data that would give guidance to database curators to decide on the inclusion of this data in the database. Also from the database user perspective a universal quality mark would be ideal.

In practice this is not so trivial. Resulting from discussions with various stakeholders in the field it was decided that EUROCarbDB should not impose strict quality constraints for the inclusion of data. We should, however, provide various quality marks with the provided data, so that end-users are able to use these as selection criteria and/or to weight the data for their own purposes.

In this report we present the quality parameters that we intend to determine and provide. In addition, we provide guidelines for optimal glycan NMR spectroscopy. It is clear that the sample amounts and the scientific questions can impose limits on the number and the types of experiments actually performed. For example, when only minute amounts are available, ^{13}C -NMR experiments are mostly not feasible.

1.2 Quality parameter of NMR evidence in EUROCarbDB

Information on sample characteristics is vital to a well organized database. The following information (when applicable) is suggested to be given for each NMR study.

1. Sample concentration
2. Solvent system
3. pH or pD
4. Salt concentration
5. Actual temperature
6. Reference used
7. Performed NMR experiments
8. Special parameters e.g. mixing time
9. Signal-to-noise ratio (S/N)

1.2.1 Sample conditions

Sample concentration has a direct link to the quality of the spectra, as the signal is proportional to the amount of sample available. It should be noted that on the other hand low sample concentrations provide narrow line-width, which is good for accurate measurements of chemical shift values and coupling constants.

The choice of solvent system, including pH and the presence of salts has considerable effect on the observable chemical shifts. Therefore it is crucial that information on the type of solvent, or solvent mixtures, is provided next to the pH or pD reading and salt concentrations.

The temperature at which the NMR experiments have been performed has a marked effect on spectra. Therefore it is a must to provide the actual temperature.

1.2.2 NMR experiments

In research projects multiple NMR experiments are likely to have been performed. Information on which experiments have been performed should be provided as well as the actual spectra. In this respect it is also important to at least provide the characteristic parameters for the experiments, such as spectrometer frequencies, mixing times, time-domain dimensions, and spectral widths.

1.2.3 NMR Processing and referencing

The processing of the NMR spectra can be performed by any suitable software package. EUROCarbDB provides such software, ProSpectND. These packages provide many options to yield processed spectra of high quality. Information relevant for the digital resolution should be provided (spectral width / number of data point) as well as the phasing parameters. In addition to the technical aspects of the processing, the choice of chemical shift reference is important. The information for spectral referencing must be provided: which reference compound / peak define which ppm value.

1.2.4 Signal to Noise

The *Signal to Noise* (S/N) ratio is an attractive qualifier. However, the amount of material available is the most important limiting factor for high S/N ratios. Whether a spectrum is good enough to provide the relevant scientific information does not heavily

rely on the signal to noise ratio, as long as the relevant signals can be distinguished from the noise. EUROCarbDB will provide S/N values and define the method for determining the S/N ratio. Consensus was reached during a targeted EUROCarbDB / CCPN workshop on the following definitions:

- 1 Noise: standard deviation of the distribution of intensities in a noise area.
- 2 Signal: one should take the height of an isolated single nucleus signal
- 3 A S/N ratio minimum of 3 seems acceptable.

1.2.5 Quality or trust-values

The chemical shift list provided is the most important piece of information that needs to be judged. EUROCarbDB partners have developed tools (e.g. CASPER) that provide chemical shift estimations for ^1H and ^{13}C chemical shifts of glycans. These estimates can not only play an important role during the assignment and spectrum analysis process. They can also be used to yield chemical shift validation information. The first application of such a validation process deals with the identification of trivial errors resulting from e.g. typos in chemical shift lists. The validation potential needs to be established in practice. However, it is not expected that the chemical shift predictions will be accurate enough to yield a flawless chemical shift validation. Nevertheless, we are confident that the shift predictions will play an important role in the curation process as well as in the quality assessment for a user that queries the database. We will present predicted chemical shifts alongside the reported values.

Chemical shift list provided

A number of quality measures can be derived from the database and automatic procedures:

- Level of completeness of the assignment: percentage ($\#$ assigned atoms / $\#$ total atoms)*100%
- Published in a peer-reviewed journal (yes/no)
- Curated (no / yes; yes: name of curator)
- Shift validation (compare provided and estimated shifts)
- Biggest outlier in the shift validation: $\text{MAX}(|\delta_{i,\text{exp}} - \delta_{i,\text{calc}}|)$
- Average difference with shift-validation $\sum(|\delta_{i,\text{exp}} - \delta_{i,\text{calc}}|) / n$

- RMSD difference for shift prediction.
- The RMSD approach is probably the best criterion that can be translated into a single number. It is broadly used in science as a measure of the goodness of fit. Its strength is that it combines the two approaches: (biggest outlier and average deviation), as the rmsd calculation combines all information into one number, but is sensitive for single outliers.
- Experiments to back up the chemical shift list.
- A scheme has been proposed to provide a trust value as a single number:
 - A shift list without any experimental information to back it up:
0 points
 - reference to a peer-review journal provided: +1
 - 1D experimental data provided: +1
 - 2D or 3D homonuclear experiments provided: +2
 - 2D or 3D heteronuclear experiments provided:
+2
 - Number of different nuclei used (and provided):
+#nuclei – 1

This results in a trust value. Here are some examples:

Plain shift list, without evidence	0
SugaBase record	1
Published shift list (ref + 1D NMR data provided)	3
Published shift list (ref + 1D and 2D 1H data)	5
Published shift lists (ref + 1D and 2D 1H and 13C data)	7
Published shift lists (ref + 1D and 2D 1H, 13C and 31P data)	8

1.3 Guidelines for good glycan NMR spectroscopy

1.3.1 Purification and desalting

NMR samples are ideally pure compounds in a well defined solvent system. Chromatography is generally the preferred method of purification. As a final purification step all proton-containing contaminating substances, such as buffer-compounds must be removed from the samples. Anionic carbohydrates and large oligosaccharides are easily desalted on small (1 cm x 20 cm) gel filtration columns eluted with water. The sample is recovered in the void volume and is ready for D₂O exchange after lyophilization. In case the sample contains high amounts of acetate, as can be the case after high pH anion exchange chromatography, the use of 5 mM NH₄HCO₃ as eluent instead of pure water facilitates the removal of the acetate. Small neutral oligosaccharides should be desalted on a mixed-bed of cation and anion exchange resins. Sample desalting can also be performed using 'Hi-trap' or carbon columns.

1.3.2 Sample preparation

Sample preparation deserves some attention with respect to the choice of solvent, to solvent conditions and to sample purity, as inclusion of protonated contaminants or solvents affect both the spectrum as the sensitivity. Depending on the type of NMR equipment various sample holders are available, ranging from conventional NMR tubes to dedicated low concentration, low volume sample holders. Glass tubes are the most common.

1.3.3 Solvent

For ^1H -NMR experiments on carbohydrates the samples are normally dissolved in deuterium oxide (D_2O or $^2\text{H}_2\text{O}$), a solvent that closely resembles the natural aqueous solution, however, without the characteristic huge amount of protons. In this way the skeleton carbohydrate protons can readily be observed. In D_2O solution all exchangeable protons ($-\text{OH}$ and $-\text{NH}$) are non-observable. Often the exchange of these protons with deuterium atoms is done in advance in one or more D_2O - dissolution/lyophilization cycles. However, in some cases one is particularly interested in these labile protons. The amide signals are needed to obtain a complete sequential assignment of a peptide moiety when present. Furthermore, the observation of carbohydrate hydroxyl signals or amine and amide resonances has definite applications in the field of conformational analysis. In these cases the sample is generally dissolved in an H_2O / D_2O mixtures with D_2O contents between 5% and 10%, needed for the deuterium lock system. It should be noted here that for the latter type of experiments the pH plays a crucial role. The chemical exchange of these labile protons is readily catalysed at a pH above 7. In many cases the pH is adjusted between pH 5.5 and pH 6.5 to ensure a slow exchange. The exchange rate of hydroxyl protons is larger than that of NH-protons. The presence of metal ions (or salts in general) will have a catalytic effect on the labile proton exchange. The occurrence of paramagnetic impurities is harmful to most NMR experiments. Metal ions are conveniently removed from the sample by passing it through a small ion exchange column.

1.3.4 Sample amounts and sensitivity

NMR spectroscopy can provide a wealth of information. However, NMR is relatively insensitive compared to other spectroscopic or spectrometric techniques. The hardware has improved with ever increasing magnetic field strengths (currently just

above 950MHz) and the availability of cryogenically cooled probe heads. Further, one should aim to introduce sufficiently large amounts of sample in the NMR instrument. This can be achieved by using high sample concentrations, and/or using NMR tubes with a large diameter. Therefore, sensitivity is not a problem in cases with ample amounts of soluble material. After several HPLC cycles absolute amounts may be extremely small (nano-mol ranges). Using the standard NMR hardware the best choice is to use special NMR tubes with a thick glass bottom and a glass insert (plunger) on top of the sample. These kind of NMR tubes are often referred to as 'Shigemi-tubes', named after the main manufacturer of this kind of tubes. The glass has magnetic properties similar to water so that the magnetic field lines are not refracted at the sample-glass interface. The result is that the effective sample volume can be reduced from 550 to 250-300 μl . Therefore, the NMR spectra can be recorded at increased sample concentrations. To reduce the sensitivity problem further, the manufacturers of NMR instruments offer 'micro-probes' or 'nano-probes'. In comparison to conventional NMR measurements, the sample volumes are drastically reduced from typically 500 μl to 150-50 μl , respectively. In this kind of NMR probes the sensitivity gain is achieved by placing the receiver coils much closer to the (smaller) sample. Especially with the nano-probe technology NMR spectra of released glycoprotein glycans can be recorded with sample amounts as small as 10 nmol.

1.3.5 Spectrum acquisition

Choice of spectrometer is not an option in many laboratories. When a laboratory does have multiple spectrometers, the scientist should consider running the spectra at higher field when sensitivity and/or signal overlap appears to be an issue. The best spectra should be provided to EUROCarbDB.

Optimization of the spectrometer prior to the experiment is a prerequisite for high-quality data. 'Tuning and matching' of the probe is like tuning your radio receiver to the proper station and should be done after each sample change, where sample content (solvent, concentration of solute, type of solute, pH and salt concentrations) has changed.

Next, the homogeneity of the magnetic field needs to be optimised. This process is referred to as shimming. Good shimming is important as the inhomogeneity of the field is directly observable as line broadening or even worse, distorted line shapes.

Shimming is essential. Proper shimming takes time and is sometimes referred to as an art rather than a skill.

Once the spectrometer is shimmed and tuned, the spectroscopist should determine the appropriate pulse-lengths and amplifier settings required for the experiment. An overview of experiments has been given in the first-year report of EUROCarbDB (available at <http://www.eurocarbdb.org/about/reports>).

Special attention should be given to optimising the digital resolution. This optimisation does not only improve the digital resolution itself and thus the accuracy at which chemical shift and/or coupling constants can be determined. It also increases the acquisition time and storage efficiency. The same S/N can be achieved in less time, or a better S/N will be achieved in the same time. Parameters such as the length of the time-domain (TD) and observable spectral width (SW).

1.3.6 Spectrum assignment

The best way to make sure that chemical shift lists are accurate and as complete as possible is to record 2D NMR spectra, or NMR spectra at even higher dimensions.

Analysis of novel glycans and/or the determination of 3D structure and dynamics of glycans require more advanced NMR techniques than 1D experiments. Signal assignment of complex oligosaccharides can partly be done by comparing occurring chemical shifts with data stored in a library of known compounds. Typically, homonuclear correlation type spectra, such as various COSY or TOCSY experiments, are needed to assign signals in the bulk area. The glycan assignment procedure generally starts from the anomeric proton signal, or any other well resolved NMR signal. The 2D spectra can be traversed from diagonal peak to diagonal peak via cross peaks. In this way the resonances in each spin system can be identified.

Homonuclear correlation experiments identify spin-systems. They are essential in this respect, but they do not provide sequence information. NOESY or ROESY spectra are mostly used for this purpose. In many cases the most intense NOESY peak identifies the linkage location. Nevertheless, examples are known where this is not the case. Here it can be of advantage to also include non-proton nuclei in the analysis such as ^{13}C , ^{15}N or ^{31}P . An unambiguous method to resolve the above issue is the HMBC experiment, which yields through-bond correlations. This yields cross peaks between

the anomeric carbon and the ring proton of the adjacent sugar residue, and between the anomeric proton and the ring carbon of the adjacent sugar residue. In order to make use of HMBC spectra the carbon resonances should be assigned as well as the proton resonances. HSQC or HMQC spectra provide one-bond ^1H - ^{13}C correlations. In other words it identifies chemical bonds between ^1H and ^{13}C , and thus provides the link between the assigned proton resonances and the carbon resonances. The HMBC experiment is basically an HMQC experiment tuned to small (1-10 Hz) heteronuclear couplings. Often the one-bond correlations are explicitly filtered out the spectrum. However, omitting this gives many one-bond correlations as well, that are convenient in the assignment process.