

2007

Bijvoet Center, Utrecht University

Participants:

ECDB-Utrecht: Bas Leeftang, Thomas Luetke, Ana Arda Freire

ECDB-Stockholm: Magnus Lundborg

ECDB-Heidelberg: Martin Frank, Siegfried Schloissnig

CCPN: Rasmus Fogh

MSD: Wim Vranken

Introduction on the meeting and its goals. (Bas Leeftang)

Limited overview of glycoscience database efforts in the past. Limited user base in general and in particular a limited user base that was willing to pay for information that could also be obtained from studying scientific journals. EuroCarbDB aims to integrate better with other life-science bioinformatics initiatives and it aims to use modern informatics tools. EuroCarbDB is being funded by the EU (FP6) as a "Research Infrastructure Design Study". Nowadays funding agencies are starting to realise that investing in preserving existing data and tools is worthwhile. So chances are better for EuroCarbDB than for CarbBank in the past. (well will see hoe this works out in a couple of years ...)

The entire NMR workflow from acquisition through processing and analysis to deposition was schematically presented. The Bijvoet Center is now offering its NMR processing package 'ProSpectND' to the EuroCarbDB team (now, and later on to the entire scientific community). ProSpectND is a user friendly package written in C and it compiles for Unix (OpenMotif widgets) or MS-windows (building from a cygwin environment under windows.) The windows binary also runs happily under WINE.

ProSpectND documentation is available from:

http://www.eurocarbdb.org/nmr/prospectnd/manual_prospectnd.html

Binaries can be obtained from this web site as well:<http://www.eurocarbdb.org/applications/nmr-tools/>

Source code will be made public later on, but is currently available upon request (or from EuroCarbDB SVN)

One special feature was demonstrated in some detail and that is the possibility to reduce data files in size. Two methods are being used here (individually or simultaneously). Two non-standard floating point data types are introduced: float*3 and float*2. Normal floating point numbers use one byte for the exponent and three bytes for the mantissa. In float*3 and float*2 we keep using one byte for the exponent (dynamical range stays unaltered), but we only use two (float*3) or one byte (float*2) for the mantissa. Effectively this means that the dynamical range stays intact, while the precision of the data is reduced. This is generally not a good idea for time-domain data. However, for fully processed (frequency domain) data, it will not have any significant effect on e.g. contour plots or cross-peak integrals.

EuroCarbDB setup (Matt Harrison)

Matt described the EurocarbDB set up. A relational database (postgres) will be used as the storage layer. The application layer will consist of JAVA code 'Data Objects' interfaced to the database through the open-source package 'hibernate'. On top of the data objects an 'Actions' layer contains all the business logic, and both Web and Soap clients work through the Actions layer rather than directly with the data objects. On the front end ECDB will use 'struts' to connect the web clients to the Actions layer. A graphical representation and extensive explanations of the code and coding methods is available (for ECDB users) on the Wiki.

(http://www.dkfz.de/spec/EUROCarbDB-Wiki/index.php/Main_Page)

As described, EurocarbDB is built as a distributed database, that caters for sharing data between different sites, as well as adding and later publishing restricted access data. It is conceived as a database rather than a LIMS system, and it is expected that data will not change much once entered. The overall structure is based on four tables connected by many-to-many links: Sequence (equivalent to CCPN Molecule), Biological context (Taxonomy, tissue, disease, ...), Evidence (NMR, MS, any exp data), and Reference. The different kinds of 'Evidence' (experimental data) are stored each in its own set of tables. With respect to NMR the most common operation will be to generate NMR project data elsewhere, and upload complete (sets of) XML files. There will also be some need to edit data directly inside the EurocarbDB database.

The CCPN data model will be integrated by slotting in the CCPN-generated database tables 'as-is' alongside the other tables in the database. The CCPN-generated Hibernate mappings and corresponding objects can then work from the CCPN tables while other Hibernate mappings take care of the rest of the data.

ChemComps (Wim Vranken, Martin Frank, Thomas Luetke)

The larger part of the first meeting day was well spent on discussion on ChemComps in general and carbohydrate ChemComps in particular. Wim Vranken introduced the CCPN/MSD concept of molecular systems, molecules, residues (ChemComps) down to atoms. Martin Frank presented the carbohydrate specific aspects relevant to ChemComps. Thomas Luetke, presented his work on setting up MonoSaccharideDB, which is/will be a valuable source for getting single unique names for monosaccharides. This system gets rid of trivial names or different names for the same molecule. As such MonosaccharideDB will be the perfect knowledge base to provide carbohydrate ChemComps, or at least the correct basetype (==unsubstituted monosaccharide).

Two different strategies were discussed. Glyco-CT describes basetype and substituents formally as two different entities. So one could easily map these entities to individual ChemComps. Alternatively, one could combine basetypes and substituents into one ChemComp.

The two approaches have different strengths:

Different entities:

- Fewer ChemComps to name, name spaces easier to handle (good).
- No problem in selecting atom names (good).
- Base type ChemComps would need large numbers of ChemCompVars so that all possible linking patterns could be accommodated (bad).

- Substituent groups would have separate residue Ids and no visible connection with the residue they were linked to, in the absence of special-purpose application code (bad).

Combined entities:

- ChemComp generation would require a user-friendly program to generate the new ChemComps, in practice from the same building blocks as the 'different entities' approach (Acceptable, such a program would probably be required anyway).
- ChemComp names would have to be generated. The names might be long in practice. There might be problems with name clashes or hard-to-understand names (Acceptable).
- Requires extra care for importing/exporting glycan sequences and translating them back and forth into ChemComps (bad).
- More consistent with the general perception of glycoscientists in what a monosaccharide or glycan residue is (good).
- Atom names would have to be generated for substituents. The agreement was to use the substituent name and number, with a suffix reflecting the substitution position. E.g. A 3-acetyl group would have carbons named C1_3 and C2_3. (Problematic. This effectively introduces a new nomenclature system for substituted carbohydrates, which might or might not fit with user habits and any existing nomenclatures).

After a fruitful discussion, it was decided that ChemComps to be used in CCPN and ECDB-NMR will be monosaccharides inclusive of substituents. This means that GlcNAc or Glc-3OMe will be represented by one ChemComp.

There is considerable overlap in MSD and EurocarbDB needs with respect to a single and consistent list of substituents. It was decided that ultimately MSD would maintain this archive of substituents, while EurocarbDB would maintain the base type ChemComps. Martin and Thomas will shortly provide Wim with the raw material (mol2 formatted files) to build ChemComps for the most urgently needed monosaccharides. Martin has a list of these.

On a separate point it was agreed that alpha, beta, and open forms of a carbohydrate, as well as different ring sizes should be handled as ChemCompVars of a single ChemComp, as these could interchange in solution. The possibility of having separate ChemComps for reducing and non-reducing versions of a carbohydrate was mooted but not agreed on. D and L sugars should be separate ChemComps.

The use of integer (artificial) residue Ids versus meaningful string residue Ids was discussed. The general opinion was that you would need both, with the data storage built on the artificial ID.

Coming features in CCPN analysis and formatconverter

Wim said that Wayne is currently implementing 1D spectra functionalities in analysis. But wanted feedback on what features are wanted. During the meeting things such as peak picking and assignment, coupling constants and preferably also identification of splitting patterns were brought up. It was decided that Magnus should compile a list of features that would be interesting and send it to Wayne.

In formatconverter it might be possible to use local dictionaries to link certain variables in a pulse

program to variable in the datamodel. This could make it possible to import e.g. mixing times from TOCSY and NOESY experiments. The dictionaries would only be valid for a certain pulse program, which means that they would have to be tailored to what is used in the respective department.

DataModel and Code-generation

Rasmus explained the current activities of reshaping the whole CCPN model and code generation stuff. It was made clear that the model and code-generating machinery used in PIMS is based on an outdated branch of CCPN. It does not make much sense to start coding ECDB stuff on the basis of the present PIMS code, although technically the set up is very similar.

Rasmus hopes to have the CCPN code base upgrade ready by the summer (Java/SQL and Python/XML API).

Rasmus will investigate whether a Java/XML API can be provided with limited efforts. This API is compatible with the present public version of FormatConverter and CCPN analysis. When this Java/XML API will be available programmers (i.e. Sigfried) can start coding the EuroCarbDB specific Java applications.

<Post meeting it looks like it will be possible to make a branch 4 Java/XML API with somewhere between days and weeks of work. As Wayne (who does the work) must go to the US in two weeks and has a lot of preparation for that as well, it is uncertain whether we can get the API ready in a couple of weeks or by the second half of April. We shall try for the early date. Rasmus>

DataModel Changes

EuroCarbDB has expressed several questions and wishes for new data to be stored in the model.

- It was agreed that Rasmus would add the possibility to store images of spectra (in DataSource).
- snRatio and snMethod will be added to the DataSource
- It is not clear what the proper location is to store mixing times, ExpSeries and/or ExpTransfer?
- <Post-meeting: There may be a better way of doing this. A separate mail is being circulated on this point. Rasmus>

Wim will supply python code to demonstrate how to get/set solvent specific stuff in/from the datamodel, without the need to implement this in Analysis. It would still be good to have a user friendly interface for this functionality.

SampleCondition can have free texts tags added. This seems to be blocked in Analysis.

<Post-meeting: In fact the area of ExpSeries, SampleConnections and mixing times would need some extra editors in Analysis if they were to be used heavily. So far no one has asked CCPN for this, so it is not yet on the TODO list. You should probably remind us of this, and tell us more about the kind of work people will want to do. Rasmus>

Some aspects of the model may give rise to problems in the longer term. These are:

- Description of partially unknown molecular topology, including incompletely known residue linking and polymers with (partially) unknown numbers of repeats.
- NMR work and assignment on molecules of not-yet-known structure.

Any information about the precise needs in these areas would be welcome.

CCPN and ECDB

It was concluded that best way to link the CCPN and ECDB data models would be to have the Evidence table pointed to by the NmrProject. The NMR-project contains the molecule and a shift list as a minimum. It can also contain an whole array of spectra that are considered to contribute to the 'evidence'. In this way one can store older shift lists (SugaBase / Glycosciences.de) in a similar manner as complete NMR projects that have been handles by CCPN Analysis. The technical set-up could be done with an intermediate class living in a EurocarbDB-specific package. Alternatively it could be 'hacked in' in a number of ways, given that only a single many-to-one link is required at the moment

Bas Leeftang, will provide example sugar shift lists to Wim Vranken. In this way Wim can see how to tweak FormatConverter to convert SugaBase entries to CCPN project files.

Bas Leeftang, Rasmus Fogh and Magnus Lundborg