

Final Report

EUROCarbDB

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

**Design Study for a New Infrastructure
implemented as a
Specific Support Action**

Contract number: RIDS Contract number 011952
Project Coordinator: Dr. Claus-Wilhelm von der Lieth (†)
Dr. William E. Hull (since Dec. 2007)
Deutsches Krebsforschungszentrum (DKFZ)
Heidelberg
Project Website: <http://www.eurocarbdb.org/>
Project Duration: 54 months from 01.04.2005 to 30.09.2009

**Project funded by the European Community
under the “Structuring the European Research Area” Specific Programme
Research Infrastructures Action**

A. ACTIVITY REPORT

Table of Contents

1. Background.....	3
2. Project Objectives.....	3
3. Participating Institutions in the Design Study.....	6
4. Summary of Project Activities	7
5. Major Achievements.....	9
6. Work Report	10
6.1. Initial Phase.....	10
6.2. Glycan Structure Encoding	12
6.3. Core Database Architecture	14
6.4. Distributed Peer-to-Peer Networking.....	16
6.5. Data Models and Standards for Experimental Evidence	17
6.6. HPLC and MS Analysis Tools in EUROCarbDB.....	19
6.7. NMR Analysis Tools in EUROCarbDB	22
6.8. References in EUROCarbDB.....	25
7. State of the Art.....	25
8. EUROCarbDB Publications	27

1. Background

Carbohydrates (glycans) and their glycoconjugates are involved in a variety of fundamental biological processes, e.g., cellular differentiation and recognition, embryonic development and fertilization. They are also involved in numerous pathological conditions such as bacterial and viral infections, inflammatory diseases, cancer and metastasis, immune response, and, therefore, offer attractive possibilities for diagnostic and pharmaceutical applications.

In comparison with proteins and nucleic acids, carbohydrates have been far less intensively studied in the past, primarily due to their structural complexity, the unique technical challenges of carbohydrate chemistry, and perhaps a lack of appreciation for their biological significance. Technical innovations during the last 30 years have transformed our capabilities for isolating carbohydrates and determining their structures in detail. These efforts have no doubt been fuelled by the ongoing discovery of numerous diverse and important biological roles for carbohydrates and their glycoconjugates.

While for the genomics and proteomics fields standardized, comprehensive, open-access databases have been established for some time, no large, uniform data collections for carbohydrates had been compiled prior to the EUROCarbDB project. The availability of such comprehensive databases, however, is a prerequisite for successful large-scale glycomics projects aiming to decipher the still largely unknown biological functions of glycans. For this purpose, common protocols and GLP¹ quality criteria for the generation of experimental data and comprehensive guidelines for the design and maintenance of databases are indispensable, especially with regard to NMR, MS and HPLC data², which are the key technologies for the identification and analysis of carbohydrates. The interpretation of HPLC profiles as well as MS and NMR spectroscopic data of complex glycans requires appropriate collections of reference data with a broad scope.

The Internet offers unique capabilities for establishing a global and interactive peer-to-peer communication system for scientific data. The development of cross-linked databases and modern informatics tools is well-recognized as an important and rapidly developing area in glycobiology/glycomics research. Two large glycomics projects have already been established. The database developed in the USA by the Consortium for Functional Glycomics (CFG), the largest glycomics project to date, can be characterized as a centralized top-to-bottom initiative to make available the data obtained within the consortium. The Japanese “glycomics in diseases” initiative (HUPO-HGPI) aims to define community standards for data representation within the field of functional glycomics in relation to diseases.

2. Project Objectives

The EUROCarbDB design study was initiated to develop a new infrastructure which will constitute the nucleus for a central depository for carbohydrate-related data (structure and function), comparable to and cross-linked with the extensively used genomics and proteomics data collections. The technical framework shall be established for a bottom-to-top initiative

¹ GLP = Good Laboratory Practice

² NMR = Nuclear Magnetic Resonance, MS = Mass Spectrometry, HPLC = High-Performance Liquid Chromatography

where all interested research groups can input their primary data and interpreted results. The initiative aims to provide systematically integrated tools for streamlining European research in glycobiology through the development of databases, bioinformatics standards, efficient analysis and search algorithms, and web-based software components. To maximize synergism, other available bioinformatics and biomedical resources will be linked to the newly created databases and cross-referenced in an efficient way.

One major aim of the EUROCarbDB design study is the development of appropriate algorithms and tools which facilitate the rapid and reliable processing, interpretation and annotation of analytical data for glycan structure determination with a high degree of automation. This effort for the glycomics field is comparable to the development of the highly automated techniques for nucleic acid sequencing which have so successfully accelerated genomics studies and enabled their large scale application. The EUROCarbDB project is expected to provide comparable impetus in glycomics.

The EUROCarbDB study has concentrated on the evaluation and development of the basic requirements for the proposed infrastructure. Four thematic areas were defined as the Design Studies DS1 - DS4 and include feasibility studies as well as essential preparatory steps for the establishment of the prototype EUROCarbDB.

- DS1 – establish standardised analytical methods and their data representations for glycoscience studies, GLP guidelines for good practice, procedures for quality control (QC); develop universal glycan structure-encoding algorithms and associated data base models.
- DS2 – design software tools which enable a peer-to-peer (P2P) network of distributed databases for the glycosciences. The availability of such a tool will encourage researchers to deposit their experimental data in a local database which may be kept private until publication. Inexpensive hardware platforms and the availability of free software tools will favour this process. Feasibility studies were also planned for the automatic cross-referencing and linking of EUROCarbDB with existing genomics and proteomics databases.
- DS3 – develop algorithms and software tools which facilitate and, where possible, automate the deposition of HPLC and MS data and the rapid and reliable analysis and interpretation of HPLC profiles and LC-MS and tandem MS/MS experiments. The existence of a sufficiently large collection of high-quality reference spectra and tools for automatic analysis is an urgent demand for high-throughput glycomics projects. An accepted central depository for the primary spectroscopic data (evidence) confirming glycan structures will considerably improve quality control efforts and reduce the loss of evidence which plagues older databases.
- DS4 – analogous to DS3, develop algorithms and software tools which facilitate and automate the deposition of NMR data and their rapid and reliable analysis and interpretation. Cooperation with the CCPN organisation³ to adapt and extend the CCPN data model, originally for NMR of proteins, to allow the use of the free CCPN software for the analysis of NMR data for carbohydrates.

To ensure the success of EUROCarbDB, it was intended that the coordinator as well as the steering committee should closely monitor the development of other glycomics database initiatives, in particular the two large projects in the USA and Japan. The existing personal

³ CCPN = Collaborative Computing Project for NMR. <http://www.ccpn.ac.uk/index.html>

interaction of EUROCarbDB participants (ICL, DKFZ) with the CFG should guarantee that each project can concentrate on its own merits and specialities with maximal synergism. The same is true for the Japanese initiative where EUROCarbDB participants (ICL, DKFZ) are members of the advisory board. The coordinator and steering committee of EUROCarbDB should also seek early contact with new initiatives to avoid redundancy of efforts and to encourage the use of the standards for structure encoding, data representation and exchange formats developed within this project.

The design goals and concepts of EUROCarbDB are far-reaching and by no means limited to the time frame or extent of the current project. The architecture of the database has been conceived to allow at a later stage further types of experimental data to be added, e.g., 3D-structures derived from X-ray studies, protein-carbohydrate binding studies, multi-dimensional NMR spectra, carbohydrate microarrays, microcalorimetry, surface plasmon resonance, or collections of methods for chemical, combinatorial, chemo-enzymatic and large-scale synthesis of glyco-probes.

Above all, EUROCarbDB has been conceived as an open-access, open-source project devoted entirely to the use of web-based or downloadable free software, whereby the users are encouraged to contribute not only data and error reports but also ideas for improvements and extensions.

3. Participating Institutions in the Design Study

No.	Organisation (name, city, country)	Short name	FoE: (fields of excellence) SR: (specific roles in the consortium)
1	Deutsches Krebsforschungszentrum, Heidelberg, Germany Project Coordinator	DKFZ	FoE: bioinformatics for glycomics, development of databases, automatic update and annotation, automatic interpretation of MS spectra. SR: project coordinator, management of the consortium, DS2 leader, involved in DS1 and DS3.
2	Universiteit Utrecht, Faculteit Scheikunde, Bijvoet Center for Biomolecular Research, Utrecht, The Netherlands	BCU	FoE: NMR-database, analysis of primary & 3D structures of glycoprotein glycans and polysaccharides, molecular interaction and structure-function relationships. SR: DS1 leader, involved in DS2 and DS4, management of the consortium webpage and virtual communications center.
3	Stockholm University, Stockholm, Sweden	SU	FoE: NMR spectroscopy, NMR prediction tools, primary sequences of carbohydrates. SR: DS4 leader, involved in DS1.
4 ^a	European Molecular Biology Laboratory - European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom	EMBL-EBI	FoE: annotation and cross-referencing with other data bases, database development and maintenance. SR: involved in DS1 and DS2, host of the central database.
5	Imperial College London, London, United Kingdom	ICL	FoE: glycobiology, glycoproteins, biological mass spectrometry, carbohydrate antigens, parasite antigens. SR: DS3 leader, involved in DS1, contact to other initiatives (CFG, HUPO-HGPI).
6 ^b	The Chancellor, Masters and Scholars of the University of Oxford, Oxford, United Kingdom	UOXF.AL	FoE: glycoproteomics, glycoimmunology, structural glycobiology, HPLC, enzymatic sequencing, biological mass spectrometry. SR: involved in DS1 and DS3.
7 ^c	Universität Basel, Basel, Switzerland	UBA	FoE: NMR spectroscopy, artificial neural networks, oligosaccharide synthesis, drug design. SR: involved in DS1 and DS4.
8	Justus-Liebig-Universität, Giessen, Germany	UGI	FoE: analysis of complex carbohydrates using separation techniques and MS. SR: involved in DS1 and DS3.
9 ^b	National Institute for Bioprocessing, Research and Training, Dublin, Ireland	NIBRT	FoE: glycoproteomics, glycoimmunology, structural glycobiology, HPLC, enzymatic sequencing, biological mass spectrometry. SR: involved in DS1 and DS3.

^a **Subcontractor to EMBL-EBI:** Dr. Tony Merry, Glycosciences Consultancy, Charlbury, OXON, OX7 3HB UK. **SR:** involved in DS1 and DS2.

^b In year 2 personnel moved from UOXF.AL to NIBRT.

^c UBA left the project after year 2.

4. Summary of Project Activities

Year 1:

- Initial setup of the project (recruitment of qualified personnel, implementation of the communications infrastructure, introduction of a reporting system).
- Development of a project homepage (www.eurocarbdb.org) and a logo.
- Setup of an external expert group.
- Broad and open discussion to establish commitments for the definitions, standards and digital formats required for all relevant data to be stored.
- Communication and exchange of information, ideas and data with other developing international glycomics projects.

Year 2:

- Final establishment of standards for the description of glycan structure and related data in light of the experience gained during the practical implementation using test data and in accordance with international agreements.
- Development, implementation and testing of data models for the core database as well as for the experimental methods (MS and HPLC).
- Development, implementation and testing of the communication network between local servers.
- Development, testing and public release of a tool for the rapid graphical generation of glycan structures for database input (*GlycanBuilder*) as well as tools to support the rapid, (semi)-automatic interpretation and assignment of MS data (*Glyco-Peakfinder*, *GlycoWorkbench*).
- Intensive cooperation with the CCPN organisation to adjust formats and descriptions of the NMR data provided by EUROCarbDB for compatibility with the CCPN data model, ensuring that both projects will work smoothly together.
- Intensive contacts with other international initiatives to adjust formats, controlled vocabularies and common interface procedures so that unhindered communication between the emerging initiatives will be encouraged.

Year 3:

- Implementation of web interfaces to upload MS and HPLC data into the central EUROCarbDB database.
- Recording and assignment of high-quality MS and NMR spectra within the framework of DS3 and DS4.
- Integration of the HPLC analysis tool *autoGU* and the corresponding database *GlycoBase* at NIBRT into the EUROCarbDB framework.
- Adaptation of the NMR analysis and structure prediction tool *CASPER* to be able to open or create CCPN projects compatible with the *CcpNmr Analysis* software.

- Addition of ^1H spectrum simulation tools to *ProSpectND* (the one- and multidimensional NMR analysis freeware developed by BCU) for the analysis of the second-order spin systems frequently encountered with carbohydrates.
- Integration into the EUROCarbDB framework of MonoSaccharideDB, which contains the standardized encoded structural descriptions of the building blocks used to create all glycan structures.
- Intensive cooperation with CCPN led to adjustment of the CCPN data model to cover the molecular structures and specific aspects of NMR analysis in the glycosciences. The newly developed CCPN API⁴ will allow smooth integration of EUROCarbDB and CCPN.
- Public beta test of the EUROCarbDB tools *Glyco-Peakfinder* and *GlycoWorkbench* for the (semi)-automatic analysis of MS data.
- Organization of workshops where people from outside the consortium were trained to use the EUROCarbDB tools.
- Publications in international peer-reviewed scientific journals: (a) methodology and format for a universal, unequivocal digital encoding format for glycan structures (*GlycoCT*) and (b) tool for graphical generation of glycan structures (*GlycanBuilder*).
- Publications in international peer-reviewed scientific journals: HPLC and MS analysis tools (*autoGU*, *Glyco-Peakfinder*, *GlycoWorkbench*).
- Further intensive contacts with other international initiatives to adjust formats and to develop controlled vocabularies and common interface procedures for communication and data exchange.

Year 4:

- Optimization of the EUROCarbDB database web-interface; development and implementation of efficient search algorithms.
- Import into the EUROCarbDB core database of ca. 13500 glycan structures (translated into GlycoCT format) and, where available, literature and database references as well as biological context information (taxonomy).
- Optimization of web interfaces to upload MS and HPLC data into the central EUROCarbDB database.
- Initial implementation of the EUROCarbDB NMR database on the basis of the CCPN data model.
- Implementation of the web interfaces for upload, display and search functionalities for NMR data.
- Integration of the NMR tools *ProSpectND* and *CASPER* into the EUROCarbDB framework; preparation of manuals and tutorials.
- Generation of a chemComp library of monosaccharides and substituents as required for the CCPN data model. Each chemComp provides an atomic description of a carbohydrate building block (residue) with connectivities and atom labels, as required for the analysis and annotation of NMR spectra.

⁴ API = Application Programming Interface

- Conversion of ca. 1300 NMR datasets from *SugaBase* into CCPN project files and upload into the EUROCarbDB NMR database.
- Further recording and assignment of high-quality MS and NMR spectra in DS3 and DS4.
- Optimization of the integration of *autoGU/GlycoBase* into the EUROCarbDB framework; preparation of manuals and tutorials.
- Further public beta testing of *autoGU*, *Glyco-Peakfinder*, and *GlycoWorkbench*; preparation of manuals and tutorials.
- Organization of workshops where people from outside the consortium were trained to use the EUROCarbDB tools.
- Development of a EUROCarbDB Wiki and documentation.
- Finalizing the EUROCarbDB framework as open-source software under the LGPL license.⁵
- Continued intensive contact with other international initiatives.
- Initiative started, in collaboration with journal editors, to use EUROCarbDB for the deposition of carbohydrate structures and evidence in the context of scientific publication.

5. Major Achievements

- Promotion of EUROCarbDB ideas and concepts in an NIH Whitepaper.⁶
- Establishment of EUROCarbDB tools and database as open-source under an LGPL license.
- Implementation of a public-access prototype EUROCarbDB database according to the guidelines developed during the design study.
- Development and implementation of GlycoCT, a novel, universal carbohydrate sequence format and encoding scheme which is now gaining international acceptance.
- Introduction of ca. 13500 carbohydrate structures in GlycoCT format into the public EUROCarbDB with cross-links to other databases.
- Incorporation of biological context data fields into EUROCarbDB (taxonomy, tissue, disease, perturbations) which utilize the internationally accepted controlled vocabularies of the NCBI and MeSH organizations of the NIH.⁷
- Availability of easy-to-use web interfaces for searching for structures and biological contexts.

⁵ LGPL = Lesser General Public License

⁶ Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics* 8 (2008) 8-20.

⁷ NCBI = National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>
MeSH = Medical Subject Headings, <http://www.ncbi.nlm.nih.gov/mesh>

- Population of the EUROCarbDB database with 1331 NMR and 131 MS datasets; link with GlycoBase at NIBRT which contains HPLC data for over 350 glycans.
- Availability of free software tools for glycan analysis by MS, HPLC and NMR, covering the entire workflow from spectrometer to database entry.
- Very successful public beta test of EUROCarbDB tools. GlycoWorkbench and autoGU are used in academic and industrial research. More than 1000 downloads of GlycoWorkbench have been registered. A leading instrument vendor, BRUKER Daltonics, is currently implementing GlycoWorkbench into their ProteinScape bioinformatics platform for storing and processing of MS data.
- Integration of the MS and HPLC tools into EUROCarbDB to streamline the direct transfer of data from the mass spectrometer to the database. Web interfaces to upload MS and HPLC experiments into the central EUROCarbDB database were optimized and tested. Prototype MS and HPLC databases are ready to be used by external test users.
- Successful collaboration between a proteomics consortium (CCPN) and our glycomics consortium (EUROCarbDB).

In general, one can state that the EUROCarbDB design study is increasingly recognized and appreciated within the worldwide community of glycoscientists, not just within the European Union. In the fields of glycobiology and glycomics our initiative has taken a leading role in the development of new, innovative bioinformatics concepts as well as providing incentives and mechanisms for standardization on an international scale. During the *EuroCarb 14* conference in 2007, the largest glycosciences conference in Europe, an entire conference session was devoted to the EUROCarbDB project. During the *EuroCarb 15* conference in 2009, a satellite workshop on glycan analysis was organized by the EUROCarbDB management, and an invited book chapter about EUROCarbDB for the *Handbook of Glycomics* has been written. The participating scientists in the project have made numerous presentations at many workshops and conferences worldwide, where keen interest in the concepts and tools of EUROCarbDB has been expressed.

Recent developments and experience clearly indicate that, following in the footsteps of genomics and proteomics, the next exploding fields in the biosciences will be glycomics and metabolomics. The EUROCarbDB initiative has removed one of the key stumbling blocks impeding progress in glycomics by providing universal standards, freely accessible analytical tools, and unique networking and data exchange capabilities. Thus, by significantly expanding data accessibility and standardization, EUROCarbDB has and will continue to have a tremendous impact on the glycosciences research community as well as on commercial biotechnology efforts in Europe and throughout the world.

6. Work Report

6.1. Initial Phase

The project officially began on April 1, 2005, and the initial phase of the project was successfully completed within the first year. All partners were able to recruit highly qualified personnel. As foreseen in the management plan, a list of appropriate advisory scientists was compiled, and they were invited to function as external experts for the project. A web-based

forum was implemented and has been active since June 2005. During the first year 82 thematic threads were opened and 438 postings were submitted from in total of 61 registered members. The domain *eurocarbdb.org* was registered for the project. A website was setup and hosted by BCU (<http://www.eurocarbdb.org>). The DKFZ was the main participant in the creation of the initial outline and content of the website. The graphical design was revised in month five when the final EUROCarbDB logo became available. The colours were adjusted to more closely match a European Union style.



A project Wiki⁸ (Figure 1) was implemented and has been used by the developers to document the software tools. The developers have also intensively used Internet-based audio communication (Skype) for detailed discussions of the design of data models, for example. Many hands-on meetings for developers took place, where tasks lists for each DS subtask and developer were assigned. Beginning in January 2006 a monthly reporting system was introduced where each developer describes the progress of the tasks assigned to his activities.

eurocarb
The EuroCarb project is concerned with the creation and maintenance of software and tools for the submission, retrieval and scientific analysis of carbohydrate structures and primary experimental data.

Project Home Downloads Wiki Issues Source

Search Current pages for Search

General

- [How to help](#)
- [Introduction for scientists](#)
- [Introduction for developers](#)

Core concepts in EurocarbDB

- [Glycan_sequence](#)
- [Biological_context](#)
- [Evidence](#)
- [Reference](#)

Development

- [Eurocarb_core_API](#)
- [Eurocarb_web-application_API](#)
- [Eurocarbdb_Webapp_architecture](#)
- [Lifecycle_of_a_http_request](#)
- [Using_nested_sets](#)

Freemarker macro libs

- [Eurocarb.lib.ftl](#)
- [TextUtils.lib.ftl](#)

« **DevelopmentBuild** Updated Dec 11, 2009 by da...@nixbioinf.org

One-sentence summary of this page.

Summary

Follow the instructions below to download, compile and start the Eurocarb portal on your local machine. There's further details regarding the individual commands within the [stable build](#) instructions page.

Prerequisites

Postgres, git-svn, JDK >=1.5, Apache TomCat

Perl modules (File::chmod, DBD, DBD::PgPP)

```
cpan File::chmod
cpan DBD::PgPP
```

Complete build guide

```
mkdir eurocarb
cd eurocarb
git svn init http://eurocarb.googlecode.com/svn/ -T trunk -b branches -t tags
git svn fetch -r 1827
git checkout -b trunk_git --track remotes/david_damerell_git_staging_area
ant setup
mkdir database/data
wget http://nixbioinf.org/eurocarb/ebi_2nd_dec_2009.sql -O database/data/ebi_2nd_dec_2009.sql
cd database/scripts/org/eurocarb/database/upgrade/
ant upgrade-ebi_2nd_dec_2009-r1804-clean
cd ../../../../../../application/Eurocarbdb
ant dist
cd ../..
export TOMCAT_BIN=/location/for/tomcat/bin/
ant start
```

Follow the instructions given, you can accept most of the default values when prompted for input. The default credentials used for the Postgres connection are, postgres and flipper, if yours are different type them in as appropriate.

Figure 1. The DevelopmentBuild Wiki webpage provides access to the Introduction for Developers web pages and various source code documentation links.

⁸ wiki = Hawaiian word for fast or quick, sometimes considered as an acronym for "what I know is..."; a term used to refer to collaborative websites for knowledge management.

6.2. Glycan Structure Encoding

All partners participated in DS1 in order to develop a consistent scheme for the representation of glycan structures, for defining biological and medical descriptors (controlled vocabularies) and various standards to be used throughout the project. The main discussion centered around the characteristics of the structural data to be stored in the database and finding the best way to encode the required information for digital storage. Complications arise because glycans are not uniform linear polymers with a small number of well-defined building blocks (residues) and linkages, as in the case of proteins and nucleic acids, but rather can be highly branched and substituted structures based on a very large repertoire of residues. It was realized early on that not only precisely defined unique structures are to be encoded but also so-called *indefinite* structures, i.e., structures for which one or more features (residue identities, linkage positions, stereochemical configurations) are not unambiguously known. Such partially defined structures, where certain structural features may remain ambiguous, are frequently obtained in the initial phase of a research study when HPLC and MS analyses can be performed with very small quantities of material. Additional analytical techniques such as NMR, which requires much larger quantities of purified material, can usually resolve the ambiguities at a later stage. The key problem was that none of the existing structure encoding schemes for glycans were capable of handling indefinite structures in a satisfactory manner. Therefore, it was decided that a new, more generalized encoding scheme (using a *residue* and *connection table* approach) must be developed to provide sufficient flexibility for covering immediate demands and foreseeable future developments. Therefore, during the second year of the project a comprehensive description and digital encoding scheme for glycan structures, called *GlycoCT*, was developed at the DKFZ. *GlycoCT* is capable of describing all structural features of complex carbohydrates, as discussed in the following documentation reports (deliverables) which are available on the EUROCarbDB website.⁹

DS1-D1: *Experimental techniques for biological data collections in glycomics and digital descriptions used for the representation of carbohydrate structures.*

DS1-SUB1: *Survey structural complexity of carbohydrates and their profiles of occurrence in various tissues, species and cells.*

DS1-D2: *Recommendations for the encoding of glycan structures and the documentation of experiments in glycomics projects. Formulation of guidelines describing the data, which should preferably enter into database.*

GlycoCT – A sequence format and namespace for complex oligo- and polysaccharides. Version 3.

DS1-D3: *Creation of a central depository for carbohydrate registry numbers.*

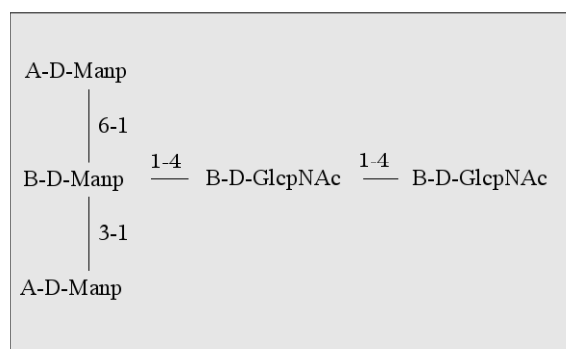
DS1-D4: *Revised recommendations for the handling of digital data produced by glycomics projects for a fully functional new infrastructure.*

An example of how *GlycoCT* describes a branched glycan in terms of a linear text code is shown in Figure 2.

⁹ All reports are or will soon be publicly available at <http://www.eurocarbdb.org/about/reports>.

In previous databases and encoding formats ambiguities or aliases for the names of monosaccharide residues are frequently encountered and severely interfere with the interpretation and comparison of structures. Therefore, the GlycoCT format also includes instructions on how to generate an unambiguous, controlled vocabulary for the monosaccharide residues used to construct glycans. The adopted procedure significantly reduces the size of the namespaces for the required carbohydrate stem types and the attached substituents. The standardized monosaccharide names and common aliases, structures, stereochemical and linkage properties and other information are stored in *MonoSaccharideDB* (see Figure 3 and Report DS1-D4), which was developed at BCU. These definitions are also used by the GLYDE-II data exchange format, which was formulated by the Complex Carbohydrate Research Center (CCRC), Athens, Georgia, USA. During the workshop *Frontiers in Glycomics* organised in September 2006 by the NIH, the Consortium for Functional Glycomics (CFG) and the Japanese Human Glycomics Proteomic Initiative, a general agreement was achieved, specifying that the GLYDE-II format will become the standard structural data exchange format for glycan databases. All of the leading bioinformatics initiatives, including many partners from EUROCarbDB, were invited and present at this conference.

IUPAC 2D-graph: N-Glycan Core



GlycoCT{HashCode}

```

RES
1b:b-dglc-hex-1-5
2s:n-acetyl
3b:b-dglc-hex-1-5
4s:n-acetyl
5b:b-dman-hex-1:5
6b:a-dman-hex-1:5
7b:a-dman-hex-1:5
LIN
1:1d(2-1)2n
2:1o(4-1)3d
3:3d(2-1)4n
4:3o(4-1)5d
5:5o(3-1)6d
6:5o(6-1)7d
  
```



Figure 2. Structure representations for a branched pentasaccharide in IUPAC format (left) and GlycoCT encoding (right), which uses a sequential (linear) text code to specify residue basetypes, substituents, and their linkages. The main RES section contains the basetypes (b) and substituents (s), both derived from a controlled vocabulary. These entities are connected via the LIN section. Residues and linkages are both numbered canonically. Linkage type identifiers allow the details of each linkage to be encoded at the atomic level.

MonoSaccharideDB by EUROCarbDB

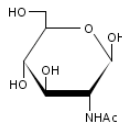
• home • notation • query • back
• overview • Monosaccharide • Substituent • Element • Help

Monosaccharide Id 1: b-dglc-HEX-1:5ll(2d:1)n-acetyl


Residue Notation Atoms Fragments All

Monosaccharide:	
Id:	1
Name:	b-dglc-HEX-1:5ll(2d:1)n-acetyl
CarbBank Name:	b-D-GlcpNAc
Composition:	O ₆ H ₁₅ C ₈ N ₁
Monoisotopic Mass:	221.1 g/mol
Avg. Mass:	221.2 g/mol
Possible Linkage Positions:	1 3 4 6 (details)

Graphical representations:

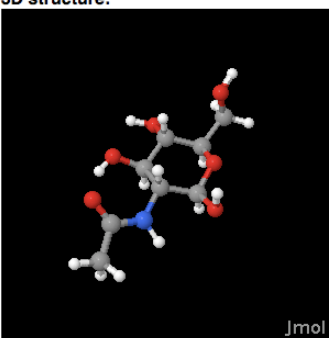


CFG Symbol:



Basetype:	
Name:	b-dglc-HEX-1:5
Size:	6
Anomeric:	beta
Abs. Config:	Dexter
Ring Type:	PYRANOSE (1 : 5)
Stereocode: ?	121220
Core Modifications: ?	none
Composition:	O ₆ H ₁₂ C ₆
Monoisotopic Mass:	180.1 g/mol
Avg. Mass:	180.2 g/mol

3D structure:



Download 3D structure file:
• PDB format

[find all monosaccharides with this basetype](#)

Figure 3. Web interface for MonoSaccharideDB, showing the information which defines one of the standard glycan building blocks *N*-acetyl-glucosamine.

6.3. Core Database Architecture

The partners involved in DS2 (DKFZ and EMBL-EBI) analysed the requirements for the intended infrastructure for the data model of the central database. Emphasis was placed on the ability to efficiently cross-link through the glycan sequence as much related biological, functional and experimental information as possible. The results of the analyses as well as the data model design are documented in the DS2-D1 report as well as on the Wiki.

DS2-D1: *Concept and implementation plan P2P-network, the central database and storage facilities.*

EUROCarbDB is based on an open-source PostgreSQL relational database. Conceptually, the database architecture (see Figure 4) comprises two major components: the *core* database and the *experimental* database. The core database, in turn, contains four subcomponents which are interrelated: *sequence*, *biological context*, *evidence*, and *references*, each of which is described in detail in the DS1-D4 Report. The **sequence database** contains the glycan structure in GlycoCT format as described above, which can be visualized in various standard graphical formats. The **biological context database** is currently subdivided into four categories: *taxonomy*, *tissue*, *disease*, and *perturbation*, whose entries are defined using standardized hierarchical dictionaries (controlled vocabularies) which are publicly available at NCBI or EMBL-EBI. An additional context called *glycoconjugates* has been prepared in the software (but is not yet activated in the public database) and is conceived to provide links to external databases for proteins and lipids, for example. Future extensions of the context fields are also possible through the use of context plug-ins. The experimental **evidence**

database contains the interpreted or annotated analytical data which confirm a given glycan structure, and these results are linked to the original raw data (see below). Finally, the **references database** contains literature references via PubMed links and cross-links to source databases from which structures were imported, for example.

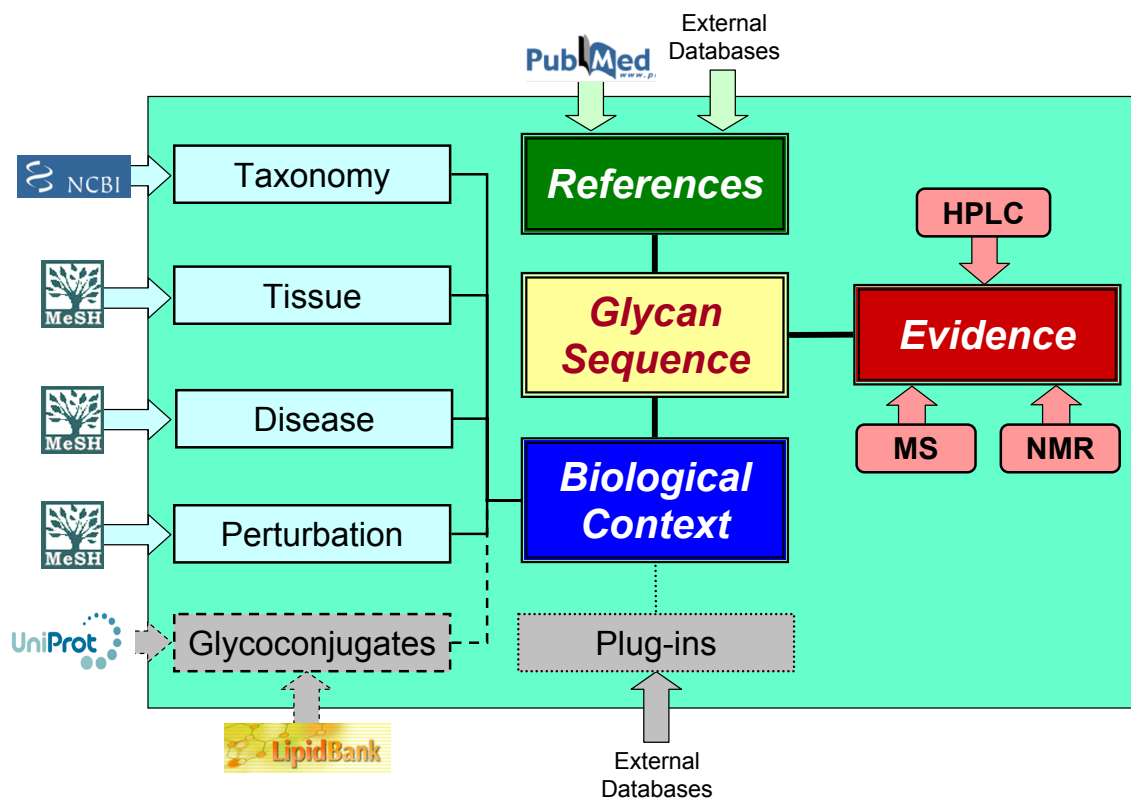


Figure 4. Architecture of EUROCarbDB. The *core database* contains the four components *sequence*, *biological context*, *evidence* and *references*. Biological context comprises several categories (light blue) which utilize the controlled vocabularies obtained from external databases (NCBI, MeSH). The dashed lines indicate that the category *glycoconjugates* has been prepared in the software but the link with UniProt has not yet been established. The dotted lines denote the capability for future extensions (plug-ins). References (green) are also obtained from external databases such as PubMed. Unique to EUROCarbDB is the internal storage of experimental evidence (red) for a glycan sequence. This information, e.g., interpreted results in the form of annotated peak lists, is derived from the raw data which comprises the *experimental database* (pink).

EUROCarbDB provides efficient functionalities for browsing and searching within the various component databases described above. For each glycan structure all associated contexts, evidences and references are listed. One can search for all glycans associated with: (a) a particular taxonomic branch or individual species, (b) a particular organ, tissue or cell type, (c) a particular disease or pathology, (d) a chemical perturbation or drug treatment. One can also search for all structures containing a defined substructure (see Figure 5). Thus, the interrelationships between component databases are of the “many-to-many” type.

The screenshot displays the EUROCarbDB web interface. The top navigation bar includes 'EurocarbDB', 'Browse', 'Search', 'Contribute', 'Tools', 'About', and 'Login'. A 'Notation' dropdown menu is set to 'CFG'. The main content area is titled 'Search Structures' and contains a search form with a 'Search' button. Below the search form, a table lists search results. The table has columns for 'Structure', 'Entered', 'Contributor', 'Data', and 'Taxonomies'. The results show four entries, each with a graphical representation of a glycan structure. The first three entries were entered on 22.04.2008 by 'maass' and are associated with 'MS' data and 'Bothrops moojeni' taxonomy. The fourth entry was entered on 21.07.2008 by 'guest' and is associated with 'Sus scrofa' taxonomy. On the left side of the interface, there is a 'Glycan structure ID' search box and a 'Sub-structure' search box. Below these is a 'GlycanBuilder' tool interface with a menu (File, Edit, Structure, View, Help) and a toolbar. The tool shows a hexasaccharide structure in Oxford graphical notation with the m/z value 1345,6723 [MONO,perMe,Na,0,freeEnd].

Structure	Entered	Contributor	Data	Taxonomies
	22.04.2008	maass	MS	Bothrops moojeni
	22.04.2008	maass	MS	Bothrops moojeni
	22.04.2008	maass	MS	Bothrops moojeni
	21.07.2008	guest	-	Sus scrofa

Figure 5. Web interface of EUROCarbDB showing the use of the *Search structures* menu. The background window shows a hexasaccharide structure element defined using the integrated GlycanBuilder tool (here in the Oxford graphical notation). The foreground window shows a portion of the results obtained following a search for all structures in the database which contain the hexasaccharide as a substructure (here shown in CFG graphical notation).

6.4. Distributed Peer-to-Peer Networking

As with other databases, EUROCarbDB was designed to store carbohydrate structures, biological context information, and database/literature references. However, a unique primary goal of EUROCarbDB from the outset of the project was the ability to capture the *raw* primary data and the annotated results which provide the experimental *evidence* for the contributed glycan structure. The management and dissemination of these data present a number of technical challenges, in terms of pure storage capacity as well as in the logistics and security implications of sharing such data across the Internet via a peer-to-peer (P2P) network. The concept was to allow users to store voluminous raw data in local distributed databases, which be defined as private (e.g., prior to publication of a structure) or public for sharing of data across the network.

Based on the concepts outlined in the DS2-D1 Report, a prototype of a P2P network was implemented connecting two nodes (DKFZ and EBI). Originally it was planned that each site participating in EUROCarbDB would have at least one node on the network. Each node in the network can have different levels of functionality, depending on the capabilities that the host institution wishes to expose. For example, data source nodes are designed to input data onto the network via the data layer, whereas sink nodes provide query facilities over the query layer (Details see). Unfortunately, significant changes in the source code of the core database were still being made on a nearly weekly so that it became impossible to maintain the complex functionality of the network for a variety of technical reasons. Therefore, the partners decided to concentrate all programming efforts in DS2 on the development of a

stable and functioning EUROCarbDB core database. It was planned that the network functionality should be then be included into the framework once the major developments on the core database were finished. However, since the key developers involved in this part of the project left during year 4, the final goal of implementing an integrated P2P network could not be achieved within the lifetime of the project.

Details of the network design are presented in the DS2-D2 Report.

DS2-D2: Implementation of P2P-network and central database.

Based on the DS1-D1 Report: “Experimental techniques for biological data collections in glycomics and digital descriptions used for the representation of carbohydrate structures” as well as the DS3-D1 report: ”Experimental and digital standards for MS spectra and HPLC profiles”

6.5. Data Models and Standards for Experimental Evidence

A key feature of EUROCarbDB is the ability to store not only glycan structures but also the experimental evidence used to determine these structures, as both raw data and interpreted and annotated results. Therefore, it was necessary in DS1 to develop standardized protocols for sample preparation, descriptions of standard experiments, data formats, result reports, etc. Furthermore, GLP and QC concepts were introduced to ensure high data quality and reproducibility. The results of these efforts are described in the DS1-D2 Report (see above) and in the detailed compilation of experiments in the DS3-D1 Report for HPLC and MS (DKFZ, ICL, UOXF.AL, UGI) and in the DS4-D1 and DS4-D2 Reports for NMR (BCU, SU, UBA).

DS3-D1: Experimental and digital standards for MS spectra and HPLC profiles.

DS4-D1: Guidelines for a uniform description of NMR spectra.

DS4-D2: Quality measures for NMR spectra.

The DS3 partners compiled comprehensive lists of all the detailed data and information for HPLC and MS experiments which are worth storing and analysed intensively the various digital standards used in MS-based proteomics projects and their suitability for glycan analysis. Database models were generated for MS and HPLC data and published on the forum for discussion among the experts. GLP and QC measures were discussed, and a questionnaire was distributed to the experts in the field. However, it soon became clear that it will be extremely difficult to develop clear guidelines and standards for the execution and reporting of glycomics experiments. Since the availability for such guidelines is essential for the field, a new working group, including members of EUROCarbDB and the CFG, was established in 2009 to address this problem. Nevertheless, data models for MS and HPLC data (ICL, UOXF.AL/NIBRT, UGI, DKFZ) were derived, implemented and tested, as reported for DS3-D2.

DS3-D2: Implementation plan of procedures for automatic inclusion of experimental data and automatic interpretation of MS-spectra and HPLC profile.

Based on the results of the task DS4-T1, *Survey of existing NMR-standards and available databases*, as reported in DS4-D1, it was concluded by the DS4 partners (BCU, SU) that the collection and assignment of high-quality NMR spectra cannot be done effectively with a Web interface alone. However, the development of a complex NMR analysis and resonance

assignment program was outside the scope of the EUROCarbDB project. An intensive search revealed that the CCPN project³ (University of Cambridge and EBI) could be an excellent partner for DS4 to overcome this problem. An initial meeting between DS4 partners and CCPN concluded that it should be possible to expand the CCPN data model (developed primarily for NMR analysis of peptides and proteins) in such a way that it can handle the complexities of branched carbohydrates as well. Therefore the steering committee decided after the first year of the project that a close cooperation with the CCPN project should be undertaken, and a productive collaboration has, in fact, been established.

This decision was logical since most of the required facilities for handling raw NMR datasets from various instrument vendors as well as procedures for the assignment and annotation of NMR resonances were already implemented in the CCPN Analysis freeware. However, some modifications were needed to adapt the software to the specific needs of carbohydrate analysis. To ensure a smooth exchange of data between EUROCarbDB and CCPN, it was decided that the CCPN data model should be adopted as provided and that only those internal tables specific for the description of carbohydrate structures will be defined and implemented in cooperation with EUROCarbDB.

Unfortunately, the CCPN team decided in year 2 of the EUROCarbDB project to implement a new optimization of their data model and the associated APIs.⁴ Thus, it made little sense for EUROCarbDB developers to begin working with the old version of the CCPN data model. Consequently, the implementation of the EUROCarbDB NMR database was delayed for more than a year. However, in the meantime, all of the tools required for general NMR data processing and spectrum simulation (ProSpectND) and for the specific analysis of carbohydrate spectra with chemical shift prediction and structure prediction algorithms (CASPER) had been developed during years 2 and 3 (BCU, SU, DKFZ).

In the final year of the EUROCarbDB project, collaboration between EUROCarbDB and CCPN was intensified and finally resulted in a working data model and NMR database, as well as a resonance assignment tool (CcpNmr Analysis¹⁰) for processed NMR spectra of carbohydrates and glycoproteins. A major task was the generation of a so-called chemComp library of monosaccharides and substituents as required for the CCPN data model. Each chemComp provides an atomic description of a carbohydrate building block (residue) with connectivities and atom labels, as required for the analysis and annotation of NMR spectra. An initial collection of chemComps for the more common monosaccharide residues was implemented in the last month of the EUROCarbDB project. CASPER now offers the facility to read CCPN projects and store prediction results as CCPN projects.

Finally, at the end of the project a test version of EUROCarbDB at BCU has been populated with more than 1300 NMR datasets which taken from the *SugaBase* collection of the CCSD¹¹ (which is no longer maintained) and converted to CCPN projects. In this way these valuable data has been again made available to the community through a modern Web interface.

¹⁰ <http://www.ccpn.ac.uk/ccpn/software/ccpnmr-analysis>

¹¹ CCSD = Complex Carbohydrate Structure Database, also called CarbBank; these data have been imported into EUROCarbDB.

6.6. HPLC and MS Analysis Tools in EUROCarbDB

EUROCarbDB is not simply a passive database system but also provides numerous free software tools to assist and automate as much as possible the complete workflow from the analytical instrumentation to the database. During years 2 and 3, the development of appropriate tools was a major goal and is one of the major accomplishments of the project partners in DS3 (ICL, UGI, UOXF.AL/NIBRT, DKFZ). These tools have been well-received and are being avidly used by the glycosciences community.

The *GlycanBuilder* (ICL) shown in Figure 6 is a rapid and flexible visual editor capable of handling complex tree-like glycan structures, using a variety of commonly used graphical symbolic notation schemes, as well as the more powerful encoding formats GlycoCT or GLYDE-II, which are difficult to produce manually. The user can rapidly specify a glycan structure by simply selecting the points of attachment of the residues, the growing structure is displayed using one of the available symbolic notations, and the output is a computer encoding of the structure in GlycoCT format. The list of structural constituents comprises an exhaustive collection of saccharides, substituents, reducing-end markers and saccharide modifications. All of the stereochemical information about a saccharide, e.g., anomeric configuration, chirality, ring configuration and linkage positions, can be specified.

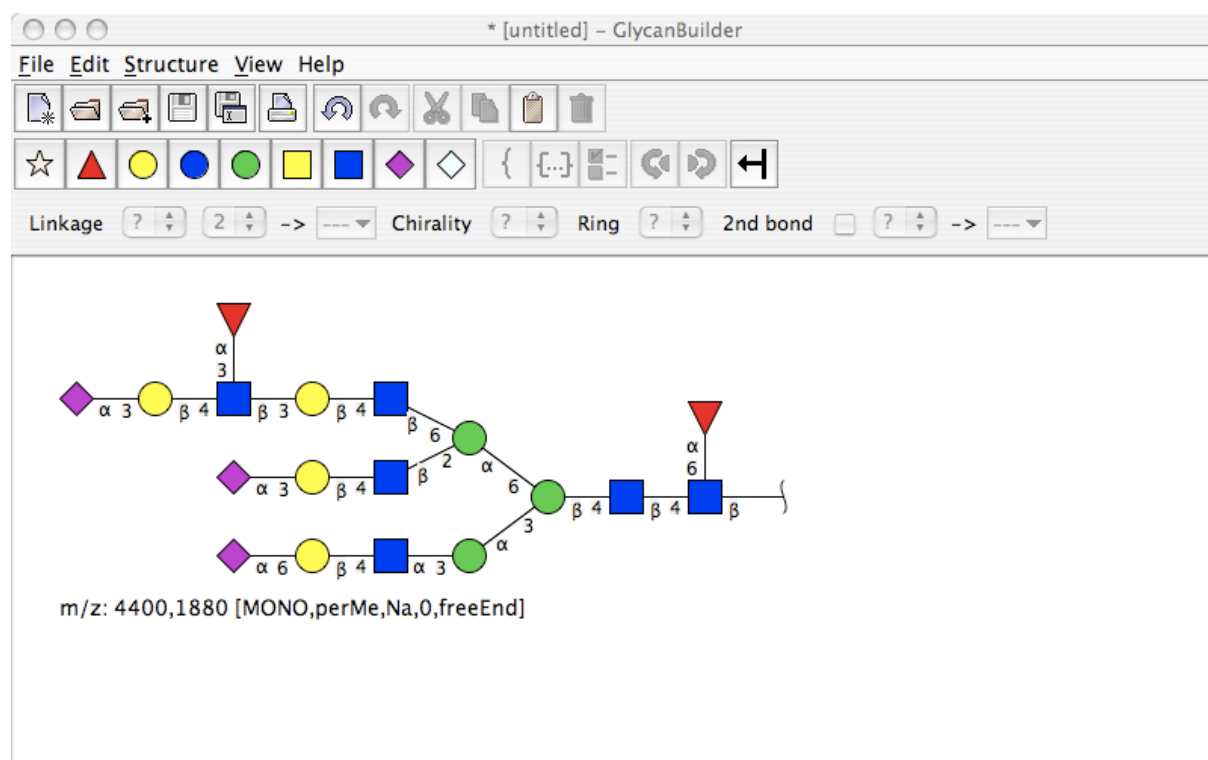


Figure 6. Symbolic representation of a glycan sequence as generated with the GlycanBuilder web tool using the CFG graphics standards.

The *Glyco-Peakfinder* (UGI, DKFZ) shown in Figure 7 was developed for *de novo* composition analysis of glycoconjugates. It is designed to ease the time-intensive manual annotation of all kinds of MS spectra. A second major field of application is the prediction and control of the annotations which will be entered into EUROCarbDB. Therefore workflows have been defined to integrate *Glyco-Peakfinder* into the input interface of the MS part of the database.

Figure 7. Web interface for the *Glyco-Peakfinder* tool in EUROCarbDB, which assists in the analysis of fragmentation patterns in mass spectrometry of glycans.

GlycoWorkbench (ICL, UGI) is a suite of software tools designed to assist the expert during the annotation of MS spectra obtained from glycan fragmentation experiments (Figure 8). The graphical interface of *GlycoWorkbench* provides an environment in which structure models can be rapidly assembled using *GlycanBuilder*, automatically matched with MSⁿ data and compared to assess the best candidate. *GlycoWorkbench* is fully integrated into the flow of data from the experimental laboratory to the EUROCarbDB database (Figure 9). The tool can be used to generate an annotated list of peaks from an MS experiment and to upload the results to the database.

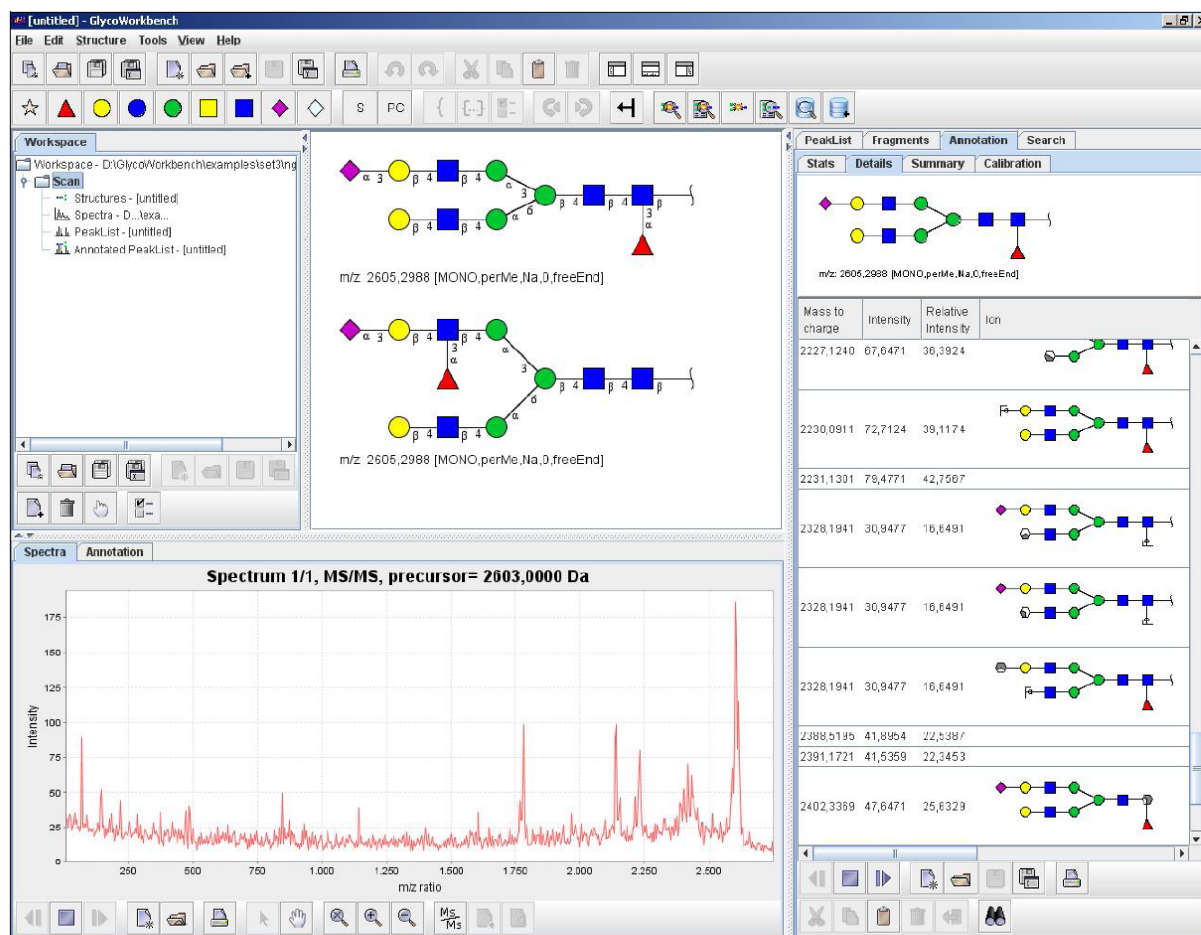


Figure 8. Web interface for the *GlycoWorkbench* tool in EUROCarbDB. *GlycoWorkbench* is an integrated suite of software tools for mass spectrometry analysis and structure elucidation of glycans and is designed to assist the annotation of glycan fragment mass spectra. All the tools are accessible from a common user interface. Here the *GlycoWorkbench* interface with the *Workspace* (left top) the *GlycanBuilder* (center top), the *Spectra viewer* (left bottom) and the *Fragmentation tool* (right) are shown. The commonly used CFG graphical symbolic notation is used to display the oligosaccharide structures.

The partners at UOXF.AL and NIBRT have developed *GlycoBase*, which is a database containing HPLC data as peak retention times in normalized glucose units (GU values) for more than 350 *N*-glycans. The web-based tool *autoGU* uses the information in *GlycoBase* to analyze and automatically annotate HPLC profiles obtained from nondigested and exoglycosidase-digested glycan samples. *GlycoBase* and *autoGU* have been installed and tested in a local version of EUROCarbDB and will soon be integrated into the public version.

The HPLC and MS analysis tools are described in detail in DS3-D2 (see above) and

DS3-D3: *Implementation of procedures for automatic inclusion of MS spectra and HPLC profiles into the P2P network.*

DS3-D5: *Analysis of data flow from spectrometer into the P2P network and evaluation of the reliability of the automatic assignment procedures for MS spectra and HPLC profiles. Recommendations of improvements for a fully functional new infrastructure.*

The software tools include Help menus and tutorials and manuals have been submitted as deliverables DS3-D4.

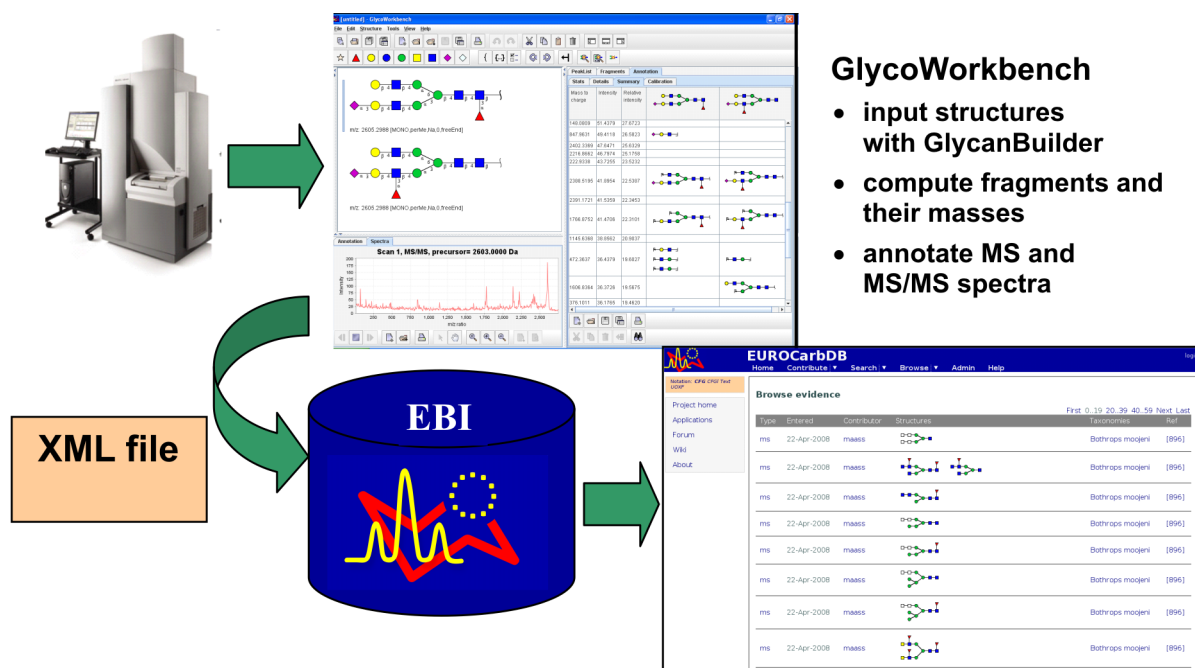


Figure 9. Schematic representation of the workflows for MS data upload in EUROCarbDB.

6.7. NMR Analysis Tools in EUROCarbDB

As described above the long-term plan for EUROCarbDB involves the integration of the CCPN data model and the open-access CcpNmr Analysis software for storing and analyzing the NMR data of glycans. Parallel to these developments BCU has further developed its downloadable multi-dimensional NMR freeware package ProSpectND¹² (Figure 10) by including 1D iterative spectrum simulation routines to assist in the analysis of the complex second-order spectra commonly observed for carbohydrates. The program is available for Windows XP, Linux, and MacOSX platforms and has been extensively used during the project to process 1D-3D NMR data obtained from Bruker and Varian instruments.

The CASPER NMR tool developed by SU is specialized for the (semi)-automatic analysis of carbohydrate spectra and can be used in two complementary ways:

- for predicting the ^1H and ^{13}C chemical shifts for oligo- and polysaccharides as well as *N*- and *O*-glycans on the basis of their primary structure,
- for predicting and ranking the likely primary structures for unknown compounds in these structure classes on the basis of measured NMR data.

Structure prediction (Figure 11) is performed by first generating all possible structures consistent with the available prior knowledge (e.g., molecular weight, monosaccharide composition, methylation analyses, MS, NMR), then predicting the chemical shifts for each structure, comparing the predictions with the available experimental NMR data and ranking the structures according to the quality of the match between experimental and predicted data.

¹² <http://sourceforge.net/projects/prospectnd/>

CASPER chemical shift predictions are empirically derived on the basis of reference data obtained under standardized conditions from a limited collection of mono-, di-, and trisaccharides. For example, for each residue in a proposed structure, chemical shifts for the corresponding monosaccharide in the reference database (or the “best match” when the correct monosaccharide is not present in the database) are taken as a starting point, and for each residue site appropriate estimates for glycosylation shift increments and vicinal interaction effects are added using reference data for the best matching di- and trisaccharide fragments in the database. Thus, the predictions are based on real measurements for substructures and not on theoretical calculations. The quality of the predictions is limited only by the limited size of the reference compound database and will improve as the database grows.

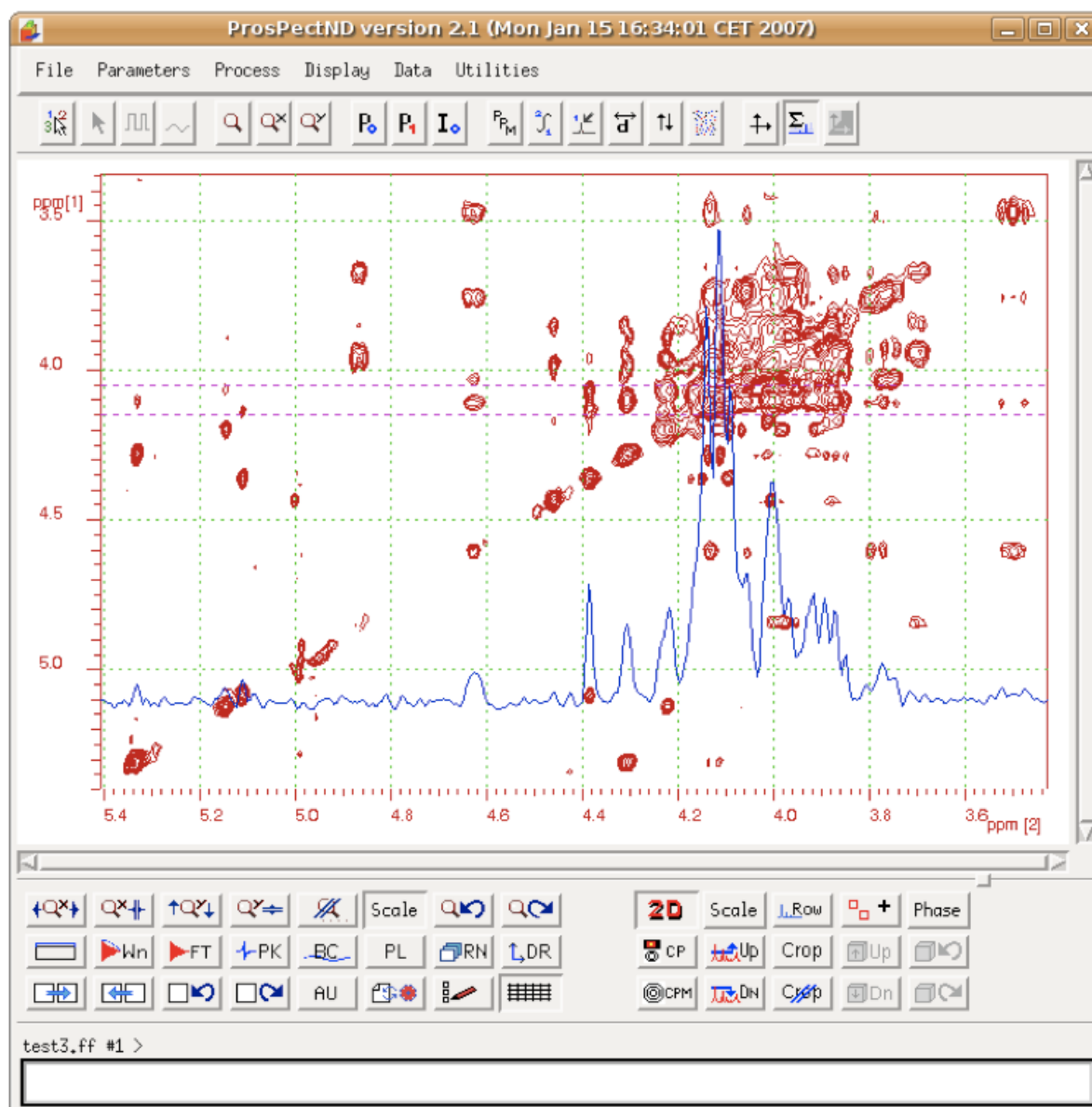


Figure 10. 2D ^1H NMR data analysis with the downloadable ProSpectND software package.

CASPER can use CCPN projects as data input and can also save results in that format. Thus, a set of experimental results (1D and 2D spectra) can be opened in the CCPN Analysis software, peak-picking performed and resonances (multiplets) defined by combining peaks, and spin systems defined with chemical shifts and coupling constants. The resulting information is then loaded into CASPER, which quickly generates structures and signal assignments which best match the experimental results.

CASPER Determine Structure

CCPN project:

Disable CCPN: No Yes

Title:

Show graphical structures: No Yes

Residue Linkage positions Chemical shifts 1D 2D

Residue	1	2	3	4	5	6	*	¹³ C chemical shifts
L-Rhap	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	101.85 79.61 70.83 73.31 70.04 17.48
L-Rhap	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	101.58 79.05 70.94 73.20 69.95 17.56
L-Rhap	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	101.85 71.47 78.28 72.45 69.88 17.36
D-GlcpNAc	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	103.03 56.39 82.24 69.40 75.21 66.97
D-Glcp	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	23.20
none	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	174.88 99.07 72.45 74.11 70.66 72.84
none	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	61.65
none	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
none	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Correct by subtracting (ppm):

Number of chemical shifts
anticipated: actual:

Minimum number of coupling constants of different magnitudes

small medium large

³J_{HH} (<2 Hz) (2-7 Hz) (>7 Hz)

Notation
CFG CFGI Text UOXF
UOXFCOL

Reference
P.-E. Jansson, R. Stenutz and G. Widmalm, Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER, *Carbohydr. Res.*, 2006, 341: 1003-10

Contact us
Have a suggestion? Want to help?
eurocarb-users@googlegroups.com
eurocarb-devel@googlegroups.com
See a problem?
report a scientific issue
report a technical problem

Figure 11. Web interface for CASPER tool in EUROCarbDB. The data input page is shown which allows the user to enter the known composition of a glycan (residues, linkages) and a set of unassigned NMR chemical shifts (in this case for ¹³C). CASPER then generates all possible glycan sequences for the residues given and predicts their NMR shifts on the basis of a large set of standardized reference data. The experimental shifts are assigned so as to give the best match with the predicted values, and a quality parameter is computed on the basis of the deviations between experiment and prediction. The ten “best” predicted structures are then presented in a ranked list, and the user can examine and compare the detailed assignments and deviations.

The NMR analysis tools are described in more detail in

DS4-D3: *Algorithms for automatic NMR spectra interpretation.*

DS4-D5: *Analysis of data flow from spectrometer into the P2P network and evaluation of the reliability of the automatic assignment procedures for NMR spectra. Recommendations of improvements for a fully functional new infrastructure.*

6.8. References in EUROCarbDB

As described above literature references represent one of the components of the EUROCarbDB core database. The available references for each glycan sequence are listed as links. A mouse click opens the corresponding reference detail page, an example of which is shown in Figure 12. The complete literature citation is given, and the *Visit the website* link accesses, for example, the PubMed entry for this publication. Furthermore, the detail page shows all glycan structures described in the current publication, and the details for each structure can be accessed by mouse click.

The screenshot shows the EUROCarbDB website interface. At the top is a navigation bar with the EUROCarbDB logo and links for Browse, Search, Contribute, Tools, and About. The main content area is titled "Reference detail" and contains the following information:

- Reference detail:**
 - Authors: Lochnit G, Geyer R;
 - Title: Carbohydrate structure analysis of batroxobin, a thrombin-like serine protease from *Bothrops moojeni* venom.
 - Citation: European journal of biochemistry / FEBS (1995) 228; 805-816
 - Link: [Visit the website](#)
- Sequences associated with this Reference:**
 - Sequence ID 13653 (link added by guest 29.06.2009)
 - Sequence ID 2284 (link added by guest 29.06.2009)
 - Sequence ID 3510 (link added by guest 29.06.2009)
 - Sequence ID 2206 (link added by guest 29.06.2009)
 - Sequence ID 13655 (link added by guest 29.06.2009)
 - Sequence ID 11915 (link added by guest 29.06.2009)
- Evidence associated to this Reference:**
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009
 - MS link added by guest 29.06.2009

Figure 12. The details page for a specific literature reference in EUROCarbDB. The literature citation is provided with a link to a database website such as PubMed, and all structures described in this publication are also shown together with links to available experimental evidence.

7. State of the Art

A public version of the functional prototype EUROCarbDB has been implemented at EMBL-EBI. The database has been filled with ca. 13500 glycan sequences in GlycoCT format, together with the associated biological context information (primarily taxonomy) and

literature references where available. The DS3 and DS4 partners have recorded and assigned many high-quality MS and NMR spectra, and these new data have been used for the development and testing of the analysis tools, the upload interfaces and the data retrieval functionalities of the database at local test installations of EUROCarbDB (BCU and SU for NMR, NIBRT for HPLC, UGI for MS). Currently 135 MS datasets are available at the public site, over 1300 NMR datasets are being tested at BCU, and the HPLC datasets of GlycoBase are accessible at NIBRT. As soon as possible the HPLC and NMR data will be migrated or linked to the public EUROCarbDB version. All of the analysis tools described above are or will be shortly available via the public website.

The field of glycomics is a young and rapidly emerging field. However, due to the inherent complexities of carbohydrate and glycoconjugate structures, the development of databases and bioinformatics tools is still lagging behind developments in the genomics and proteomics areas. To catch up, it is important that the existing initiatives in glycomics communicate with each other – regarding databases as well as applications. Therefore, the EUROCarbDB partners have intensively pursued contacts with the other international projects and initiatives, emphasizing the need to install accepted standards for data formats and data exchange. The EUROCarbDB concepts and tools have been presented and promoted at numerous international meetings throughout the project lifetime.

In summary, one can certainly stipulate that the EUROCarbDB project has already received a high degree of international visibility and support within the glycoscience community and has been well-received by the other large international projects such as CFG and HUPO-HGPI. In many respects EUROCarbDB had taken the leadership in organizing the exchange of concepts and ideas for the further concerted development of glycoscience database applications and associated bioinformatics tools.

The NIH workshop *Frontiers in Glycomics*, which can be regarded as a key milestone for the development of informatics in this area, formulated *three recommendations of high priority*:

- *Develop a robust, centralized database of curated glycan structures.*

It was agreed that the significant progress made in the area of glycan structure databases and structure encoding algorithms within the EUROCarbDB project make it the logical site to host the unified glycan structure database for the future. Further development of this resource will facilitate the integration of diverse data collections (such as MS and NMR data) housed at numerous institutions.

- *Develop an infrastructure to implement a worldwide network of databases containing experimental and analytical data relevant to the structures and functions of glycans.*

This is essentially the network concept which the EUROCarbDB design study has begun to implement (see the DS2-D2 Report).

- *Support the development of open source software for automated analysis of analytical data and data mining in the glycomics domain.*

With the expert tools developed under DS3 and DS4 for the (semi)-automatic interpretation of experimental data, the EUROCarbDB project has made a tremendous contribution in this direction and set standards for further progress in this area.

Finally, the published NIH White Paper⁶ includes significant contributions from partners of EUROCarbDB who participated in the various NIH focus groups.

In the last year of the project (2009), a proposal for mandatory carbohydrate structure deposition in the context of scientific publication (analogous to procedures in proteomics and

genomics) was prepared by the EUROCarbDB management and distributed to the editors-in-chief of the journals *Glycobiology*, *Carbohydrate Research*, *Glycoconjugate Journal* and *Journal of Carbohydrate Chemistry*. A special session during the final EUROCarbDB meeting in Dublin (Sept. 2009) was dedicated to the possible role of EUROCarbDB in this context. The editors-in-chief of *Glycobiology* and *Glycoconjugate Journal* were present at the meeting and the use of EUROCarbDB as a framework for depositing glycan sequences and their associated supplementary materials (analytical data) were discussed. In principle, the editors were supportive of this concept and were particularly interested in a system that would provide carbohydrate registry numbers, which is a feature of EUROCarbDB. However, active commitment of the journals will require that EUROCarbDB progress from a design study to a fully developed and established infrastructure, similar to the Protein Databank, UniProtKB, or GenBank. Therefore, further funding and leadership will be required.

8. EUROCarbDB Publications

The following publications describe EUROCarbDB concepts and tools and their applications.

- Abd Hamid UM, Royle L, Saldova R, Radcliffe CM, Harvey DJ, Storr SJ, Pardo M, Antrobus R, Chapman CJ, Zitzmann N, Robertson JF, Dwek RA, Rudd PM. **A strategy to reveal potential glycan markers from serum glycoproteins associated with breast cancer progression.** *Glycobiology* 18 (2008) 1105-1118. [Epub 2008 Sep 25].
- Artemenko N, Campbell MP, Rudd PM. **GlycoExtractor - a web-based interface for high throughput processing of HPLC-glycan data.** *J. Proteome Res.* (in press).
- Bleckmann C, Geyer H, Lieberoth A, Splittstoesser F, Liu Y, Feizi T, Schachner M, Kleene R, Reinhold V, Geyer R. **O-Glycosylation pattern of CD24 from mouse brain.** *Biol. Chem.* 390 (2009) 627-645.
- Bleckmann C, Geyer H, Reinhold V, Lieberoth A, Schachner M, Kleene R, Geyer R. **Glycomic analysis of N-linked carbohydrate epitopes from CD24 of mouse brain.** *J. Proteome Res.* 8 (2009) 567-582.
- Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM. **GlycoBase and autoGU: tools for HPLC-based glycan analysis.** *Bioinformatics* 24 (2008) 1214-1216.
- Ceroni A, Dell A, Haslam SM: **The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures.** *Source Code Biol Med* 2007, 2:3.
- Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM: **GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans.** *J Proteome Res* 7 (2008) 1650-1659.
- Cosgrave EFJ, Struwe WB, Campbell MP, Kattla JJ, Rudd PM. (2010). **Glycomics.** In: *Comprehensive Biotechnology*, ed. M Butler: Elsevier (in press).
- Harrison MJ *et al.* **EUROCarbDB: an open-access platform for glycoinformatics.** (2010) in preparation.
- Harvey DJ, Merry AH, Royle L, Campbell MP, Dwek RA, Rudd PM. **Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds.** *Proteomics* 9 (2009) 3796-3801.
- Haslam SM, Julien S, Burchell JM, Monk CR, Ceroni A, Garden OA, Dell A. **Characterizing the glycome of the mammalian immune system.** *Immunol. Cell Biol.* 86 (2008) 564-573.
- Herget S, Ranzinger R, Maass K, von der Lieth C-W: **GlycoCT-a unifying sequence format for carbohydrates.** *Carbohydr Res* 343 (2008) 2162-2171.
- Jansson PE, Stenutz R, Widmalm G. **Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER.** *Carbohydr. Res.* 341 (2006) 1003-1010.
- Maass K, Ceroni A: **Software Tools for Semi-automatic Interpretation of Mass Spectra of Glycans.** In: *Bioinformatics for Glycobiology and Glycomics: An Introduction.* C.W. von der Lieth, T. Lütkeke, M. Frank (eds.): Wiley-Blackwell (2009) pp. 257-268.

- Maass K, Ranzinger R, Geyer H, von der Lieth C-W, Geyer R: **"Glyco-Peakfinder" - de novo composition analysis of glycoconjugates.** *Proteomics* 7 (2007) 4435-4444.
- Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS. **Frontiers in glycomics: bioinformatics and biomarkers in disease.** An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics* 8 (2008) 8-20.
- Ranzinger R, Herget S, Lutteke T, Frank M: **Carbohydrate Structure Databases.** In: *Handbook of Glycomics.* R.D. Cummings, J.M. Pierce (eds.): Elsevier (2009).
- Saldova R, Wormald MR, Dwek RA, Rudd PM. **Glycosylation changes on serum glycoproteins in ovarian cancer may contribute to disease pathogenesis.** *Dis. Markers* 25 (2008) 219-232. Review.
- Stenutz R: **Automatic Spectrum Interpretation Based on Increment Rules: CASPER.** In: *Bioinformatics for Glycobiology and Glycomics: An Introduction.* C.W. von der Lieth, T. Lütteke, M. Frank (eds.): Wiley-Blackwell (2009) pp. 311-320.
- Tissot B, Ceroni A, Powell AK, Morris HR, Yates EA, Turnbull JE, Gallagher JT, Dell A, Haslam SM. **Software tool for the structural determination of glycosaminoglycans by mass spectrometry.** *Anal. Chem.* 80 (2008) 9204-9212.
- Tissot B, North SJ, Ceroni A, Pang PC, Panico M, Rosati F, Capone A, Haslam SM, Dell A, Morris HR. **Glycoproteomics: past, present and future.** *FEBS Lett.* 583 (2009) 1728-1735.
- von Witzendorff D, Maass K, Pich A, Ebeling S, Kölle S, Kochel C, Ekhlesi-Hundrieser M, Geyer H, Geyer R, Töpfer-Petersen E. **Characterisation of the acidic N-linked glycans of the zona pelucida of prepuberal pigs by a mass spectrometric approach.** *Carbohydr. Res.* 344 (2009) 1541-1549.

<p style="text-align: center;">B. FINAL MANAGEMENT REPORT (FINANCIAL INFORMATION)</p>
--

The following separate documents are provided.

Appendix 4 - Summary financial report for the entire project.

Appendix 5 - Human Effort Table.

<p style="text-align: center;">C. FINAL REPORT ON THE DISTRIBUTION OF THE COMMUNITY FINANCIAL CONTRIBUTION</p>

Within 60 days after receipt of the final payment by the Commission, the Coordinator will submit

Appendix 6 - Final report on the distribution of the Community's contribution.

D. QUESTIONNAIRES

The following reporting questionnaires, covering the entire project duration, will be completed and submitted on-line by March 19, 2010.

D.1 Final science and society reporting questionnaire (Appendix 1)

to be submitted by the co-ordinator only.

D.2 Final reporting questionnaire on workforce statistics (Appendix 2)

to be submitted by each contractor.

D.3 Final socio-economic reporting questionnaire (Appendix 3)

to be submitted by each contractor.