

**Contract for a DESIGN STUDY Implemented as a SPECIFIC
SUPPORT ACTION**

EuroCarbDB

<http://www.eurocarbodb.org/>

**Design Studies related to the development of distributed, Web-
based European Carbohydrate Data Bases**

Life Sciences, Genomics and Biotechnology for Health

RIDS Contract number 011952

Design Study DS1: Definition of standards, rules and formats for the biological and analytical data to be collected. Definition of good practice, recommendations of procedures for quality control.

Task Title: "Digital representation for special carbohydrate"

Deliverable DS1-T5: Report: "Creation of a central depository for carbohydrate registry numbers"

Dissemination: PU

Partners: 1,4

Due date of deliverable: 31.03.2008

Actual submission date: 31.03.2008

Start date of project: 01.03.2006 **Duration:** 48 months

Organisation name of lead contractor for this deliverable:

Deutsches Krebsforschungszentrum, Heidelberg, Germany

Author Dr. Matt Harrision , Dr. Kim Henrick

Task DS1: Recommendations for standards, digital formats and quality measures

Objectives

- Definition of a minimum description of glycomics experiments
- Guidelines for good practice and quality control
- Convention for encoding metadata such as taxonomy, physicochemical properties, biological functions
- Definition of a broadly accepted representation of carbohydrate structures
- Creation of a system of registry numbers for carbohydrates
- Maintaining EUROCarbDB the virtual communications centre

Description of DS1-T5

It is well known from genomics, proteomics as well as chemometrics, that the creation of a unique key to describe a certain molecule is the most efficient way to unambiguously identify a certain compound and to link various databases. Based on the structural description to encode glycan structures as defined in task DS1-T3, DKFZ, and EMBL-EBI will develop algorithms and procedures, which generate a unique key for each glycan structure. These procedures will require the topology of the glycan structure as often used publications as sole input. A central repository for carbohydrate registry numbers will be made publicly available at DKFZ as well as at EMBL-EBI. Everybody through the Internet can automatically access the registry number. An internet forum will enable an intensive exchange of experiences made by external users using the glycan registry tool. In such a way, the information associated with a given glycan structure can be unambiguously linked between various databases. In addition, the classification of the glycan (N-, O-glycan having a certain number of antenna) will be provided as well as a list of known biological occurrence of the requested glycan.

DS1-T5 Report – Implementation of a central depository for carbohydrate registry numbers

Considerations for the design of a carbohydrate registry numbering system

An important consideration of any modern sequence database/depository is addressing the question of what constitutes a unique entry in the depository, and how a unique entry is globally unambiguously identified. For carbohydrates, the question of what is a unique sequence is non-trivial, due to the fact that submitted carbohydrate sequences may not be fully characterised. For example, structures may be contributed that have (a) uncertain

and/or unknown internal connectivity between residues, (b) one or more unknown inter-residue glycosidic linkage elements, (c) the specific identity of one or more monosaccharide residues may not be known, and/or (d) structures may contain repeat regions. These are termed indefinite structures or indefinite sequences, as opposed to definite structures/sequences, in which all structural elements are fully described. The distinction between definite and indefinite sequences is important for informatic purposes; certain workflow paths and algorithms are not feasible for indefinite structures, and consequently necessitate special handling or manipulation, for example, discrete mass is not calculable for certain types of indefinite structures.

The screenshot shows the EUROCarbDB website interface. At the top, there is a blue header with the EUROCarbDB logo and navigation links: Home, Contribute, Search, Browse, Admin, and Help. A user is logged in as 'matt'. On the left, there is a sidebar menu with links: Project home, Applications, Forum, Wiki, and About. The main content area is titled 'Glycan sequence detail' and shows a glycan structure diagram. Below the diagram is a table with the following information:

Contributor	matt
Date Entered	Nov 29, 2007
Evidence for this sequence	ms entered on Nov 29, 2007 by matt ms entered on Nov 29, 2007 by matt
Biological contexts in which this sequence has been observed	(There is no biological context information for this sequence)
References	There are no references associated to this structure.

EuroCarbDB is a Research Infrastructure Design Study Funded by the 6th Research Framework Program of the European Union (Contract: RIDS Contract number 011952)

Figure 1 The User View of a Glycan sequence

Since the majority of contemporary methods for carbohydrate structural elucidation are incapable of completely determining sequence/structure without additional data and/or methods being performed, the ability to capture these indefinite structures as well as fully-determined, definite structures was an important consideration in the EuroCarbDB DS1 structure and standards design process. This is reflected in the capability of the GlycoCT format to encode virtually all known forms of carbohydrate structural uncertainty that may be contributed to EuroCarbDB by scientific researchers.

At a certain level of structural uncertainty however, it becomes pragmatic, and more useful from an information technology viewpoint, to consider some sources of structural

uncertainty in indefinite structures as a set of multiple, more definite structures, and to encode other sources of uncertainty into a single sequence. Each of the specific forms of carbohydrate structural uncertainty are individually described below.

Structures with indefinite connectivity: Connectivity in this context, refers to the existence of a covalent linkage between 2 monosaccharide residues, as distinct from an unknown linkage, in which a linkage is known to exist between 2 residues in a sugar, but the specifics of linkage position and/or anomeric configuration are not known.

While the GlycoCT format is potentially able to encode structures containing unknown connectivity into a single sequence, EuroCarbDB has adopted the approach that structures of uncertain or unknown connectivity are stored within EuroCarbDB as multiple structures of definite connectivity in the database. This is based on the rationale that the incidence of structures of indefinite connectivity submitted to carbohydrate databases is generally low, and that the number of definite structures resulting from the expansion of connectivity indefinite structures into multiple structures of definite connectivity is relatively small, typically less than 10. It is also often the case that the contributing researcher may have additional information that narrows the number of possible structures that satisfy their observed data.

For example, a researcher may contribute a structure in which he/she can demonstrate the presence of 2 sub-branches that may be attached to 3 possible positions of a common core structure, but that only structures with sub-branches attached to positions 1 and 3 or positions 2 and 3 but not positions 1 and 2 are supported by the available evidence. This case is simply handled by in EuroCarbDB by the introduction of the 2 sequences that satisfy the researcher's evidence as individual, definite structures.

Ensuring that indefinite carbohydrate sequences with unique connectivity are represented as unique sequence entries in the database has several very important advantages over trying to encode multiple possible connectivities into a single sequence entry:

1. It is cleaner and more intuitive informatically - multiple alternative sequences/structures are stored in the database separately, allowing them to be stored separately and referenced individually, since these are, in reality, separate structures. In this manner, other researchers are able to contribute evidence for or against the various sequence alternatives

by directly linking to individual sequence records.

2. The ability to reference multiple alternative sequences/structures individually means that alternative sequences can be ranked or scored. This is not possible if structures with different connectivity are encoded into a single sequence as a single database entry.
3. The number of alternative structures from an experiment is immediately obvious (from the number of structures) without needing to parse an all-in-one sequence.
4. Encoding structures with uncertain connectivity into a single sequence precludes the use of certain graph traversal and searching algorithms in the database that assume or require unique connectivity.

Structures with indefinite linkages: As with all types of carbohydrate sequence uncertainty, an indefinite structure may be considered as a multiplicity (mathematical set) of definite structures, however, in the case of linkage position and anomeric configuration uncertainty, the number of definite alternatives to structures with even just a few uncertain linkage positions is often prohibitively large.

Accordingly, structures that are indefinite because of unknown or uncertain linkage position(s) and/or anomericity are stored as individual, unique entries in EuroCarbDB, and handled within the encoding of the GlycoCT sequence format. That is, structures that have identical connectivity but comprise a different set of unknown linkage positions and/or anomericity are also introduced to EuroCarbDB as different sequence entries. Structures that have identical connectivity and an identical set of unknown linkage positions and/or anomericity are effectively regarded as the same sequence entry, that is, the researcher's respective data and evidence are linked to the same sequence in the database.

Structures with indefinite monosaccharide identity: Large numbers of carbohydrate structures may also be contributed in which the precise identities of one or more monosaccharide residues are not known. For example, in mass spectrometry, it is the general case that monosaccharide identity can only be determined to the level of isobaric (equal mass), so-called *generic* monosaccharides, for example hexoses (Hex), N-acetylhexosamines (HexNAc), deoxy-hexoses (dHex). As with sequences that contain indefinite glycosidic linkage elements, structures containing multiple indefinite monosaccharides are handled within the GlycoCT format and are accordingly stored within the database as single entries.

Structures with repeat regions: Certain types of carbohydrate structures are comprised of multiple repeats of one or more sub-structures, the number of which may also be unknown or uncertain. In EuroCarbDB, structures with an indefinite number of sub-structure repeats are stored as individual, unique sequence entries in the database.

The carbohydrate sequence registry

As reported elsewhere, EuroCarbDB is first and foremost a distributed database. The ability to identify a specific structure both within and without the project is of course crucial. In the EuroCarbDB network, the EuroCarbDB node located at the European Bioinformatics Institute, Hinxton is always regarded as the “master node” for the purposes of assigning a permanent, unique identifier to each and every carbohydrate sequence submitted to the project. The EuroCarbDB network in turn handles the dissemination of new sequences and their assigned ids throughout the network.

EuroCarbDB carbohydrate sequence identifier generation

Actual carbohydrate sequence ids will be generated from the underlying relational database of the EBI master node, using an established unique integer generation mechanism. The generated ids will be purely synthetic in the sense that the id is generated completely independently of the features of the deposited carbohydrate sequence. In computing science terminology, we have opted for a [surrogate key](#), rather than a [natural key](#), primarily because generating natural keys for a carbohydrate sequence that is guaranteed to be globally unique is a difficult and labour-intensive task, and secondly, synthetically generated ids are generally regarded as being more flexible and robust in an information technology sense.

As discussed above, generated carbohydrate sequence ids and sequences will be disseminated from the primary (EBI) node to all connected client nodes by the EuroCarbDB peer-to-peer network. Procedures have been designed to silently resolve the case where a client node has internally assigned a sequence id that conflicts with the official EBI-mandated id.

EuroCarbDB entry identifier generation

In addition to the generation of ids that are specific for carbohydrate sequences, EuroCarbDB provides an additional external id, whose purpose is to specifically identify a EuroCarbDB *entry*. An entry is defined by EuroCarbDB as the conjunction of a unique

carbohydrate sequence and a unique biological source (or “biological context” in EuroCarbDB parlance), from which the given carbohydrate sequence has been identified. The concept of a “biological context” in EuroCarbDB has been reported elsewhere, but succinctly, it is an abstract entity that unifies all biological source information about the biological origins of a carbohydrate structure or sequence.

So whereas a EuroCarbDB carbohydrate sequence id will identify a specific carbohydrate sequence, irrespective of biological origin, a EuroCarbDB entry id will identify a specific carbohydrate sequence from a specific biological origin only. The entry id itself is generated from the concatenation of a carbohydrate sequence id and a EuroCarbDB biological context id, separated by a hyphen ('-'). Defined in this way, a carbohydrate sequence id is trivially derivable from an entry id. It is left to the scientific community and to other biological database developers linking to EuroCarbDB to determine which of the 2 described EuroCarbDB ids is most appropriate for their purposes.

Accessibility of ids

Both the glycan sequence id and EuroCarbDB entry id will be readily available for any/all carbohydrate structures displayed in the EuroCarbDB web interface, as well as all associated tools in which glycan structures are displayed. Furthermore, carbohydrate structures and entries are readily queryable by id, as well as by sequence/sub-structure, taxonomy, tissue, mass, composition, reference and various other criteria.

EUROCarbDB login

Home Contribute Search Browse Admin Help

Project home
Applications
Forum
Wiki
About

Search Eurocarbdb

[Search by biological source](#)
including: taxonomy and tissue, as well as diseases and chemical perturbations of biological sources.

[Search by glycan mass](#)
including: searches for a specific monoisotopic or average mass, as well as discrete mass ranges.

[Search by composition](#)
including: presence/absence of specific monosaccharides, as well as monosaccharide ranges.

Enter a taxonomy/species name:

Enter a tissue name:

Enter a disease name:

EuroCarbDB is a Research Infrastructure Design Study Funded by the 6th Research Framework Program of the European Union (Contract: RIDS Contract number 011952)

Figure 2 Search Interface

The screenshot shows the EUROCarbDB website interface. The top navigation bar includes the logo, 'EUROCarbDB', and links for Home, Contribute, Search, Browse, Admin, and Help. A 'login' link is in the top right. A left sidebar contains links for Project home, Applications, Forum, Wiki, and About. The main content area displays taxonomy data for 'Eukaryota' under the path 'Taxonomy > cellular organisms > Eukaryota'. It lists the rank as superkingdom, NCBI Taxonomy id as 2759, and the direct parent taxon as cellular organisms, with 145 glycan structures encompassed. Below this, it lists direct sub-taxa: Acanthamoebidae, Acantharea, Alveolata, Apusomonadidae, and Arcellinida. A section titled 'Glycan sequence distribution within sub-taxa' lists various organisms including Bos taurus, Equus caballus, Festuca rubra, Gentiana lutea, Glycine max, Homo sapiens, Japanese monkeys, Leishmania major, Leishmania mexicana, Leucojum, Mus musculus, Oryctolagus cuniculus, Phytophthora megasperma, Sus scrofa, Trypanosoma cruzi, Tylopilus felleus, and Ulmus glabra. The word 'Vertebrata' is prominently displayed at the bottom of the list. A footer note states: 'EuroCarbDB is a Research Infrastructure Design Study Funded by the 6th Research Framework Program of the European Union (Contract: RIDS Contract number 011952)'.

Figure 3 Display of Taxonomy data

The screenshot shows the EUROCarbDB website interface for a user named 'matt'. The top navigation bar includes the logo, 'EUROCarbDB', and links for Home, Contribute, Search, Browse, Admin, and Help. A 'logout matt' link is in the top right. A left sidebar contains links for Project home, Applications, Forum, Wiki, and About. The main content area displays a personalized message: 'User home for contributor matt' and 'You have contributed to 10 sequence entries.' Below this, there are several glycan structure diagrams. A section titled 'Experiments' contains a table with the following data:

Experiment name	Date added	Experiment steps
matt's experiment	Nov 23, 2007	Mass spectrometry Mass spectrometry

A footer note states: 'EuroCarbDB is a Research Infrastructure Design Study Funded by the 6th Research Framework Program of the European Union (Contract: RIDS Contract number 011952)'.



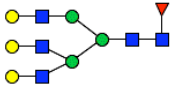
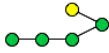





Figure 4 Customised User interface Feature

EUROCarbDB login

Home **Contribute** Search Browse Admin Help

Project home
 Applications
 Forum
 Wiki
 About

Browse Structures

Structure	Entered	Contributor	Data	Taxonomies	Ref
	Nov 22, 2007	Carbbank	-	Vertebrata Homo sapiens	[69] [784] [231] [297] [70] [788]
	Nov 22, 2007	Carbbank	-	Vertebrata Homo sapiens	[69] [784] [231] [297] [70] [788]
	Nov 22, 2007	Carbbank	-	Sus scrofa	[825] [811]
	Nov 22, 2007	Carbbank	-	root	[106] [107]
	Nov 22, 2007	Carbbank	-	root	[108] [109]
	Nov 22, 2007	Carbbank	-	root	[108] [110]
	Nov 22, 2007	Carbbank	-	Homo sapiens	[120] [122]
	Nov 22, 2007	Carbbank	-	Homo sapiens	[120] [123]
	Nov 22, 2007	Carbbank	-	Homo sapiens	[120] [136] [759] [829] [828] [733] [124] [137]

Done Adblock

Figure 5 Browsing Glycan structures in the database