Glycomics and Glycoproteomics Databases in Japan and Asia

Hisashi Narimatsu

Research Center for Medical Glycoscience (RCMG), National Institute of Advanced Industrial Science and Technology (AIST), Central 2, 1 – 1-1, Umezono, Tsukuba, Ibaraki, 305 – 8568, Japan.

E-Mail: h.narimatsu@aist.go.jp

Received: 22nd January 2014 / Published: 22nd December 2014

Abstract

The "Integrated Database Project" was initiated to establish a publically accessible database to integrate all useful life science databases in Japan. Our JCGGDB (Japan Consortium for Glycobiology and Glycotechnology Database) was selected as a promotion program in the project, focussing on the integration of all the glycan-related databases and establishment of user-friendly search systems. As part of the project, we intend the integration of databases not only within Asia but also with other countries. Working closely with various institutes in Japan and the world, we continuously develop base technologies for the database integration, facilitate interactions between databases in the field of glycoscience as well as other associated study areas, and build bioinformatics tools to support experimental study. Our goal is to create a truly useful database that could be easily and intuitively understood by every user.

Introduction

Since 2001, several glycoscience projects have been initiated by the New Energy and Industrial Technology Development Organization (NEDO) in Japan. AIST RCMG has been playing the central role in these projects. The first project was named the "Glycogene Project (GG-P)" and was undertaken to comprehensively identify the human glycosyltransferase genes and analyse their substrate specificities. The second project called the

This article is part of the Proceedings of the Beilstein Glyco-Bioinformatics Symposium 2013.
www.proceedings.beilstein-symposia.org

"Structural Glycomics Project (SG-P)" was conducted to develop technologies to analyse the structures of glycans. On the basis of the knowledge acquired and the technologies developed via GG-P and SG-P, we successfully conducted the subsequent "Medical Glycomics Project (MG-P)" focusing on the functional analysis and clinical applications of glycans.

In GG-P, we used the latest bioinformatics techniques and comprehensively searched for the glycosyltransferase genes in the database for human genome sequencing, which was almost completed at that time. The obtained data were registered in the in-house database to be used for our ongoing research work.

In 2007, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan initiated the "Integrated Database Project". This project was designed to establish a publically accessible database that would integrate all the diverse life science databases in Japan. Our glycoscience databases were also included in the project. We started publicising some of our databases and collaborated with other researchers from the field of glycoscience in Japan. At the same time, we changed the name of our database to Japan Consortium for Glycobiology and Glycotechnology Database (JCGGDB). The institutes that collaborated to form JCBBDB are: AIST, Noguchi Institute, Ritsumeikan University, Soka University, and Riken. This project was taken over from MEXT by the Japan Science and Technology Agency (JST). Although the grant from JST for this project will expire in March 2014, a subsequent project will be conducted thereafter (Figure 1).

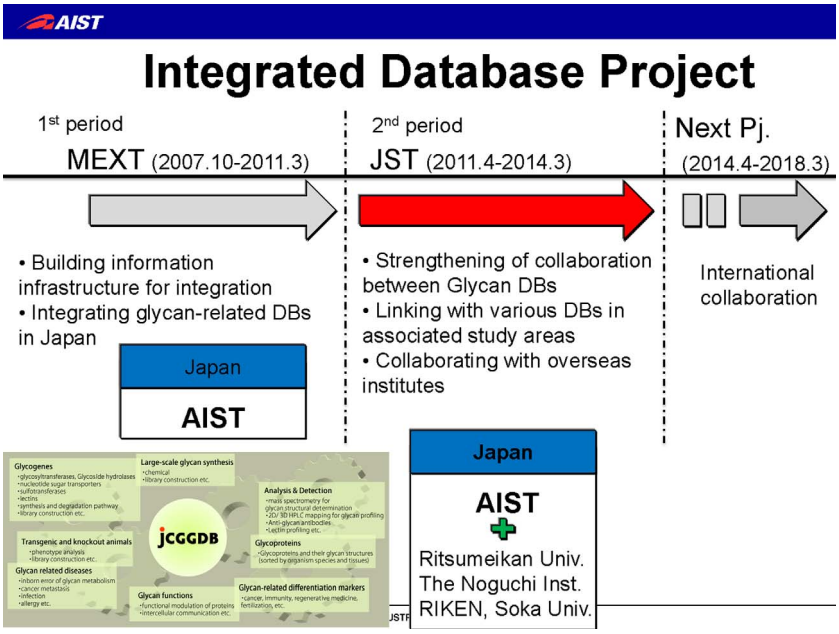

Figure 1. Schedule and objectives of the Integrated Database Project.

Results

Figure 2 shows all the databases we have developed so far. Of these databases, the Glycogene Database (GGDB), Lectin Frontier Database (LfDB), Glycan Mass Spectral Database (GMDB), and Glycoprotein Database (GlycoProtDB) contain the experimental data generated by AIST (Table 1).

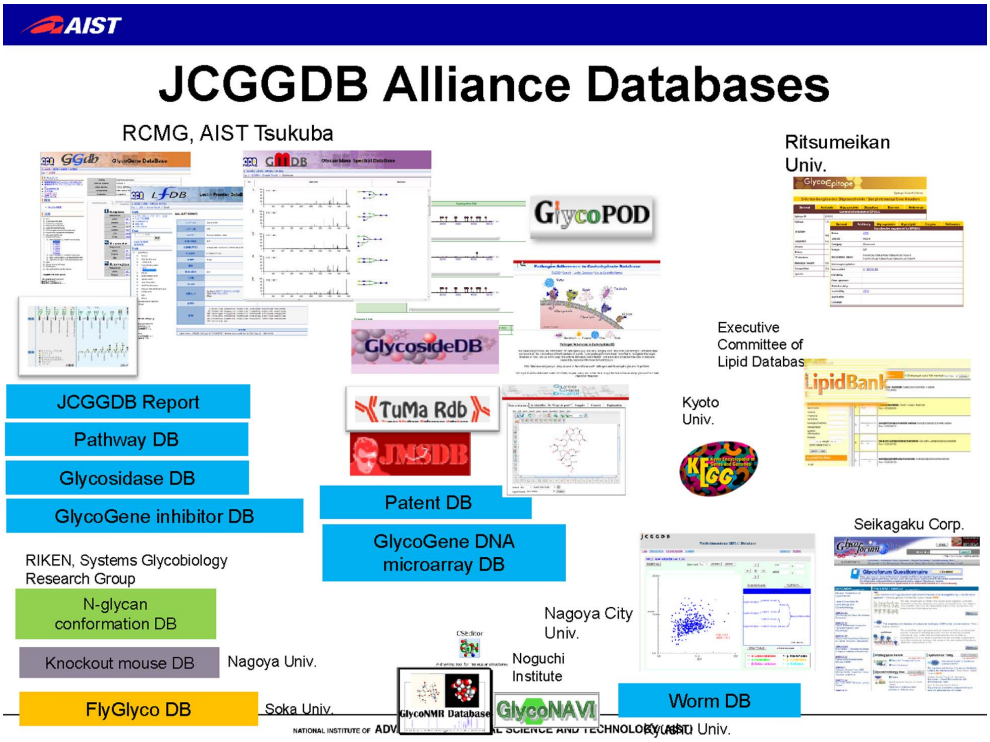

Figure 2. Current JCGGDB Alliance Databases.

Table 1. JCGGDB Database Alliance in Japan.

AIST Databases	JCGGDB Databases
GlycoProtDB	GlycoPOD
LfDB	GlycoEpitope
GMDB	GlycoNAVI/GlycoNMR
GGDB	Glycan Annotation DBs

1. GGDB (with almost 200 entries, Figure 3) provides information on the substrate specificity of glycosyltransferases, sugar-nucleotide synthases, and sugar-nucleotide transporters via the gene-based entries.

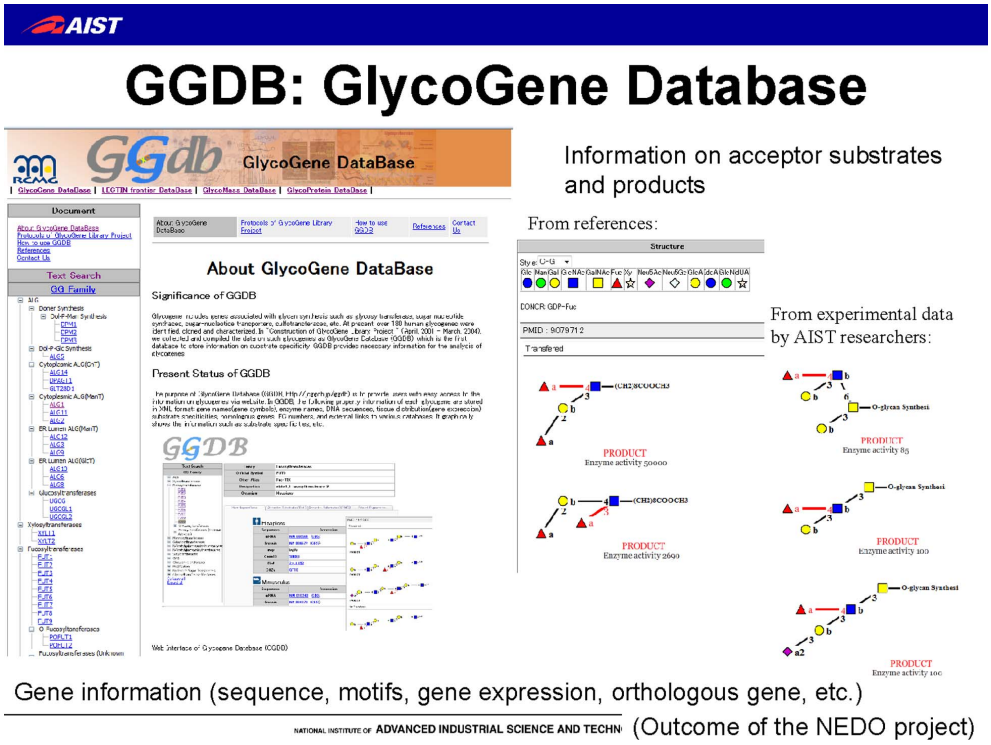

Figure 3. GGDB: GlycoGene Database.

2. GGDB is an outcome of our first NEDO project, GG-P, which originally contained only the known glycosyltransferase genes and their homologs sequences obtained through the initial human genome sequencing for in-house use during daily research. Information on other genes was subsequently collected from external databases and research papers to establish a publically accessible database. Many similar databases with information on genes are currently available on the web. Therefore, we are trying to extensively revise this database by focusing on the enzymatic and biological relevance of the glycogenes.

3. LfDB (with almost 80 entries, Figure 4) provides affinity constants of lectins and glycans measured using frontal affinity chromatography. More than 80 lectins immobilised onto the columns were tested using more than 100 referential glycan compounds. The number of the entries for the affinity data is continuously growing.

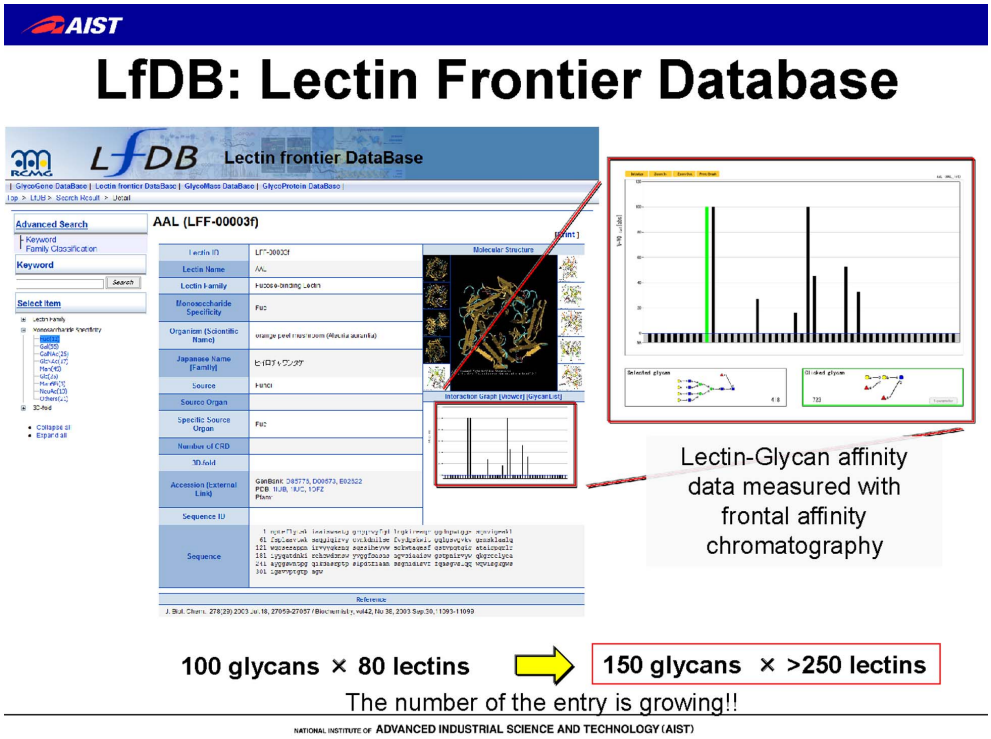

Figure 4. LfDB: Lectin Frontier Database.

4. GMDB (with almost 3000 entries, Figure 5) provides comparable mass spectrometry (MS) profiles of isomeric glycan molecules. The data were obtained from as many glycan structures as possible, including those of the commercially acquired referential compounds; in cases where no commercial referential compounds were available, we synthesised the compounds in our lab by enzymatic or organic chemical synthesis. GMDB contains the *m*/*z* values of tandem-MS obtained by performing matrix-assisted laser desorption ionisation (MALDI)-MS, which enables the structural analysis including identification of isomers.

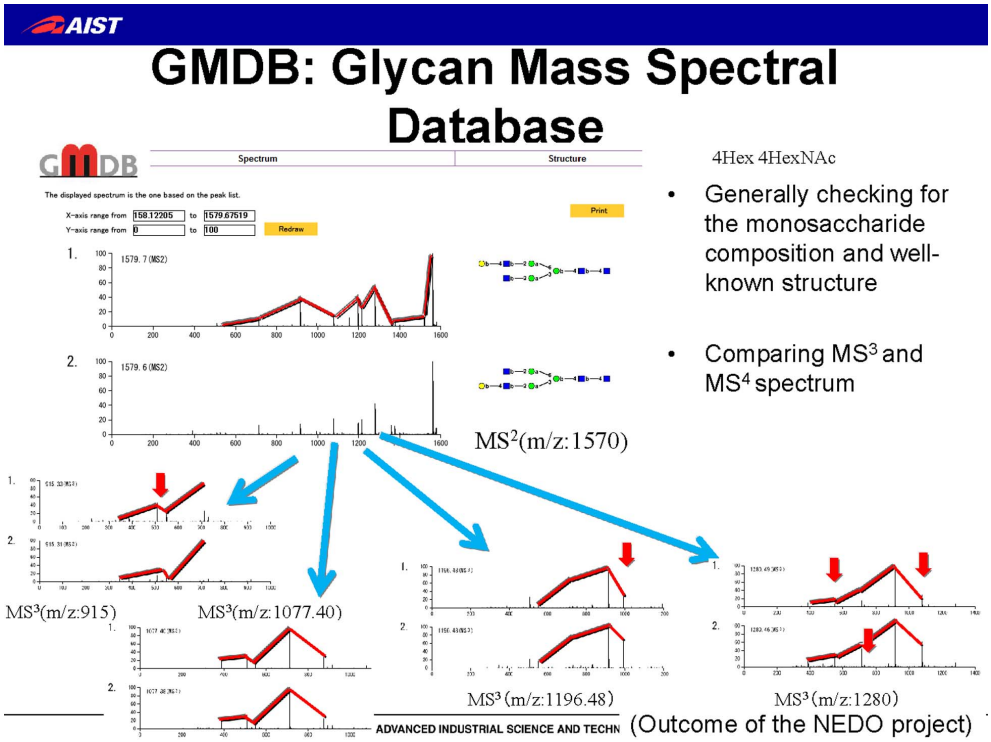

Figure 5. GMDB: Glycan Mass Spectral Database.

5. GlycoProtDB (with almost 3000 entries, Figure 6) contains data on the attachment sites of *N*-glycans obtained experimentally through the proteomic analysis of *N*-glycosylated glycoproteins in humans, mouse, nematode, and drosophila.

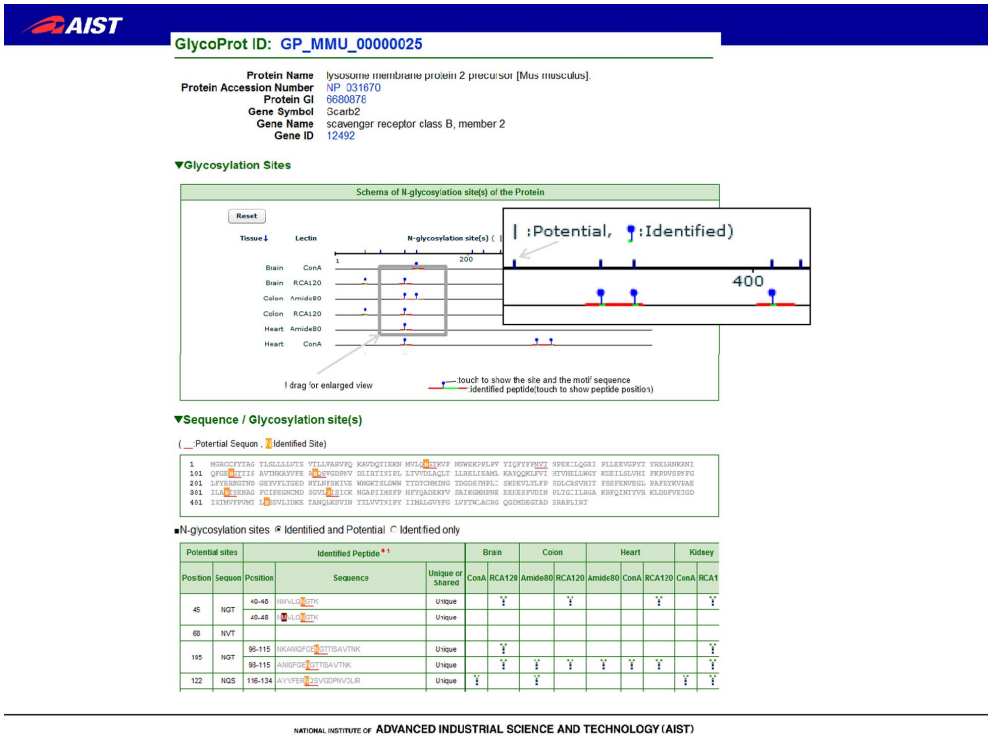

Figure 6. GlycoProt DB: Glycoprotein Database.

The principle technology is based on the LC/MS shotgun protein identification method, but we selectively enrich and concentrate the glycopeptides from crude samples by using affinity columns with lectin probes. Then *N*-linked glycans on the glycopeptides are enzymatically isolated using stable-isotope-labelled water ($H_2{}^{18}O$) and are labelled with the isotope. During this reaction, asparagines isolated from the *N*-linked glycans are converted into aspartic acid containing ^{18}O. This technology, named "lectin-IGOT-LC/MS" (LC/MS spectrometry combined with lectin-mediated affinity capture and isotope-coded glycosylation with site-specific tagging), improves the reliability of the MS analysis. The obtained amino acid sequences of the glycopeptides are mapped with respect to the complete protein sequences; then the matched results are shown by the GlycoProtDB. The pins shown on the amino acid sequences in the figure indicate the active glycosylation sites of *N*-linked glycans identified by the IGOT method. The glycopeptide enrichment required in this method also provides information on the lectin specificity of *N*-glycans, which would be very useful for determining *N*-glycan structures if combined with the data in the LfDB.

The GlycoProtDB also contains metadata on tissues and lectin columns used for sample preparation. Thus, this database also provides the data on glycoproteins (glycoforms) specific to the original tissues.

The databases not derived from the experimental data of AIST

Other than the four databases mentioned above, we also offer access to many other useful databases (Figure 2 and Table 1), e. g.,

1. Glycan Structure Database (with almost 30000 entries) houses the glycan structures collected from the literature.
2. Monosaccharide Database (with almost 800 entries) contains the chemical structural formula of monosaccharides.
3. Pathogen Adherence to Carbohydrate Database (PACDB, with almost 1700 entries) provides information on pathogens (e. g., bacteria, fungi, and viruses and their toxins) adhering to the carbohydrates expressed on the cell surface of host animals or plants.
4. Tumour Markers Reference Database (TuMaR DB, with 82 markers and 438 case studies) houses information on tumour markers and applicable cases as well as their specificity and sensitivity, as reported mainly from the clinical site.
5. Glyco-Disease Genes Database (GDGDB, with 80 diseases) contains information on hereditary diseases caused by mutations in glycan-related genes.

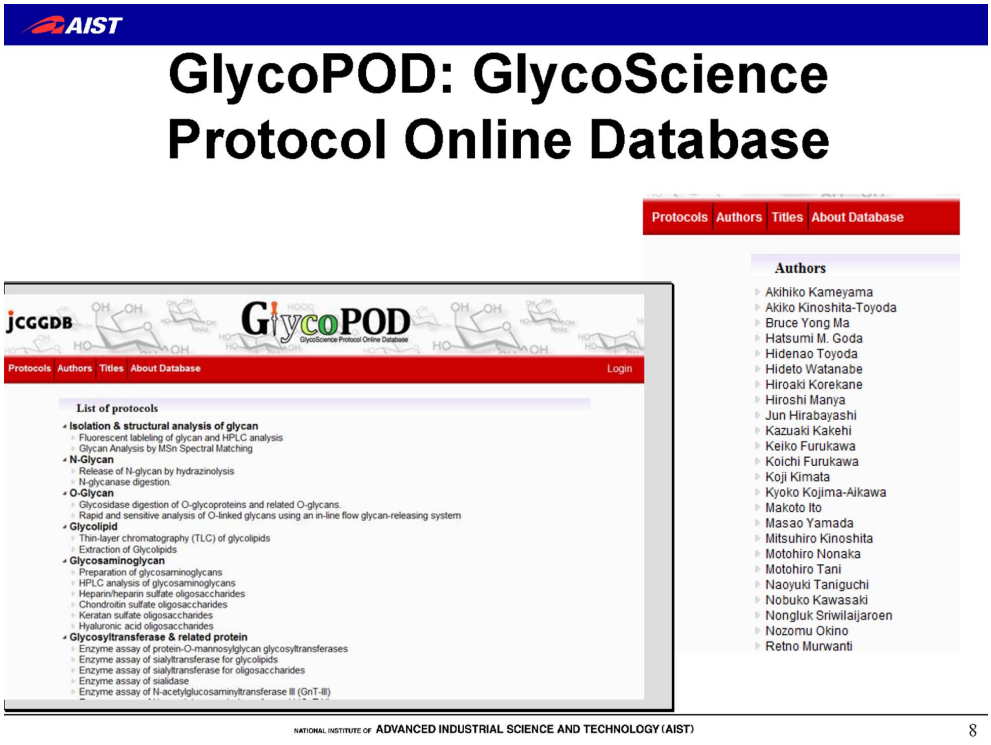

Figure 7. GlycoPOD: GlycoScience Protocol Online Database.

6. Glycoscience Protocol Online Database (GlycoPOD, with almost 200 entries, Figure 7) provides comprehensive experimental methods for studying glycoscience, including protocols for purification, analysis, fractionation, synthesis, and functional analysis.
7. Glycoside Database (Glycoside DB, with almost 70000 entries) contains information on the molecular structures of glycosides, which contain a sugar bound to another functional group via a glycosidic bond.
8. Glyco Epitope Database (GlycoEpitope, with 578 entries, Figure 8) covers information on the characteristics, usage, and applications of glyco-epitopes and antibodies.

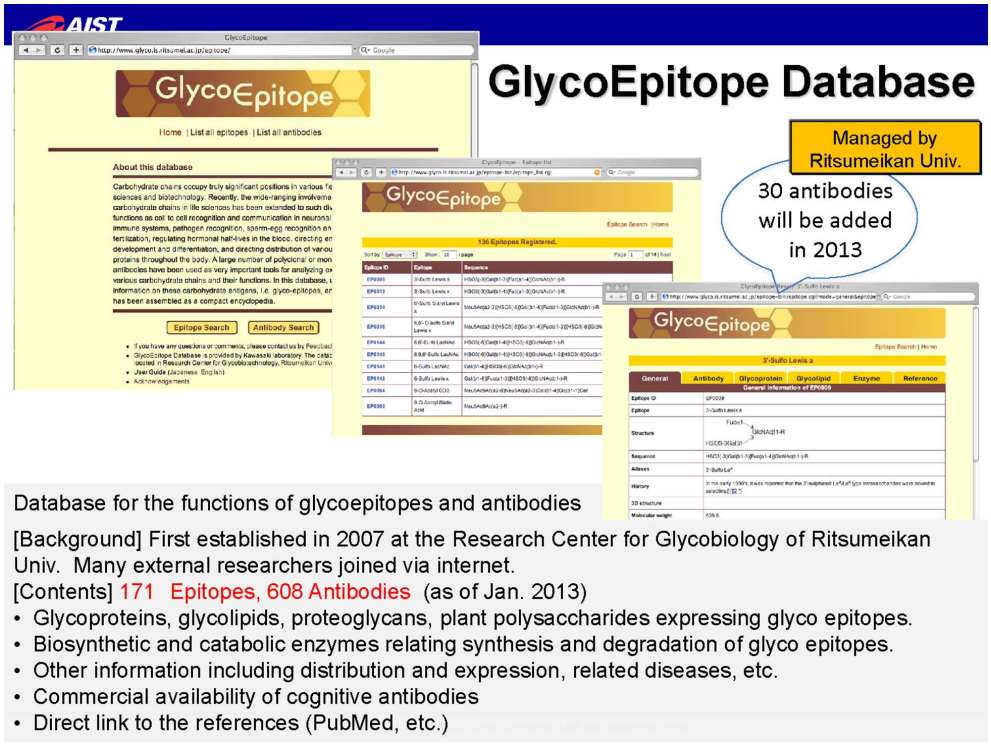

Figure 8. GlycoEpitope Database.

9. Glyco Navigation System (GlycoNavi, data on 3219 reactions) is a comprehensive database of synthetic methodology in organic chemistry, as well as information on the reagents, reaction conditions, and related literature, as well as authentic glycans and the NMR spectra of the synthetic intermediates.
10. FlyGlycoDB (with 89 entries) is a database on the phenotype of glycan-related genes in *Drosophila*, with knockdown glycan-related genes.

Mutual collaboration with the research community and data providers

The established JCGGDB was recognised as the official database of the Japan Consortium for Glycobiology and Glycotechnology, and we could build a cooperative framework with the researchers of glycoscience to integrate related databases. Moreover, we interacted extensively with the Asian researchers of glycoscience through the Asian Communications of Glycobiology and Glycotechnology (ACGG) and held regular meetings for establishing the ACGG database (Figure 9). Aiming for international standardisation, we collaborated with the world's leading databases (Figure 10), namely, UniCarbKB (Australia), GlycomeDB (Germany, US), Bacterial Carbohydrate Structure Database (BCSDB, Russia), and MonosaccharideDB (Germany), to build a glycoscience ontology in the Resource Description Framework (RDF). Simultaneously, we developed a standard notation for the glycan structure called the *Web3 Unique Representation of Carbohydrate Structures for the Semantic Web* (WURCS), and released version 1.0. We have been hosting hackathons and meetings with researchers from the US, Australia, Germany, Russia, and other countries at least once a year. We also invited researchers to the hackathons held by the Database Center for Life Science of Japan (DBCLS) for promoting collaborative research, which accelerated the development of glycoscience ontology and RDF-standardisation for each database. Thanks to the standardisation with RDF, the identifications of glycan structures are mutually linked to the data in the other standardised glycan-related databases, thus establishing the linkage between the glycan structures and information on all related proteins and lipids. This invites a broad spectrum of researchers from the fields outside glycoscience, such as molecular biology, biochemistry, and proteomics, as potential users of the repository, and thus, the number of future users may reach several millions worldwide.

Figure 9. Asian Communications of Glycobiology and Glycotechnology.

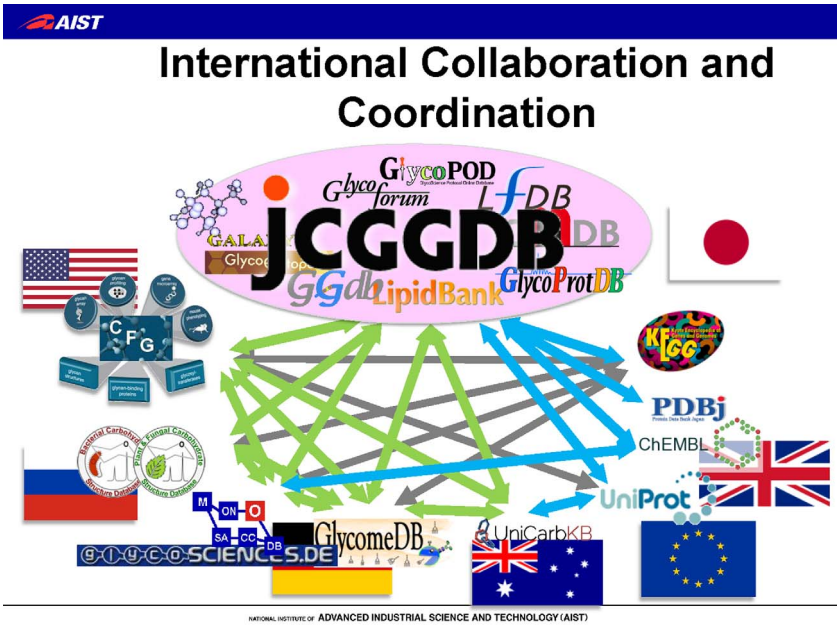

Figure 10. International Collaboration and Coordination.

Conclusion

Foreseeable tasks involved in the development of individual databases

GGDB: The analysis of glycogenes is almost complete. GGDB is currently revised accordingly.

LfDB: The contents should be increasingly enhanced. Active search for lectins naturally occurring in plants or fungi should result in the enhancement of the database. Important areas of research include modification of specifications in recombinant lectins and the analysis of lectins derived from animals, including humans.

GMDB: As there are many similar databases actively developed in other countries, collaboration and integration with such databases will be important.

GlycoProtDB: This database needs to include rapid updates reflecting the latest technical advancement. However, there has been no technology for top-down structural analysis of glycopeptides present at a small amount. In addition, it is not possible for the current technology to analyse a slight amount of a specific glycopeptide within a biological sample containing abundant non-specific molecules. A new MS-based technology to address this concern is awaited.

We have developed the lectin-microarray technique, which enabled large-scale lectin profiling of biological samples, such as various tissue specimens prepared by microdissection, cultured cell lines and their supernatants, blood, and other body fluids. Our experiments are producing large quantities of data every day; these will become available through the database in the near future.

The long-term goal in the development of glycan-related databases is the popularisation and promotion of glycoscience by increasing the recognition of glycans at a level similar to that of the gene symbols. The knowledge on glycoscience is to be collected and organised that the data can be used by researchers from the broad fields of life sciences. The technical target is knowledge sharing between glycoscience and other life science fields through a common platform, i. e., the semantic web technology. To achieve the integration of the glycoscience databases with those from other research fields via standardisation, the important base technology will be formed by standardisation of the glycan structure data and development of the international repository system. Along with the development of the base technology, we will further enhance the international coordination founded through past activities and projects.