2

Necessity to Promote Glycoinformatics

SHUICHI TSUJI *(Affiliated Fellow)* AND JUNKO SHIMADA
Life Science and Medical Research Unit

1 Introduction

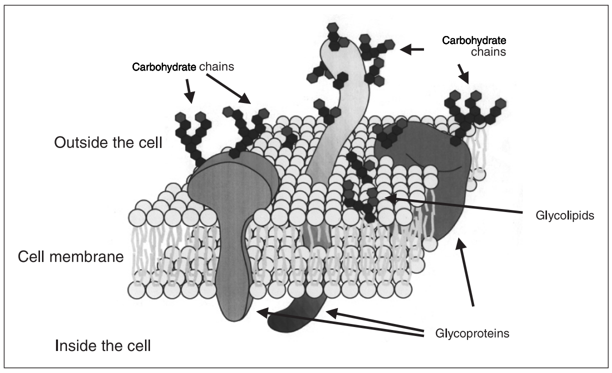
Sugar as biological energy resources, such as glycogen, starch and cellulose, and components of cells and tissues has been widely known. Yet, there is another type of sugar: carbohydrate chains existing mainly on the cell surface and bound to proteins and lipids. Recently, the carbohydrate chain is being focused on as the third chain molecule following nucleic acid and protein. Trends in carbohydrate chain research was reported in "Sugar chains as the Third Biomolecule, and Post-genome Research"[1] in the fourth issue of Science & Technology Trends-Quarterly Review. Since this report was published, research projects have been launched promoting carbohydrate chain studies. This time, we will discuss that, in order to obtain abundant fruits from carbohydrate chain research, we need to boost not only the current projects for promoting carbohydrate chain studies but also glycoinformatics, a new field of study for accelerating functional analysis of the carbohydrate chain at the molecular level.

2 Structural features and roles of the carbohydrate chain [1-4]

2.1 Structural features

Most proteins and lipids of eukaryotic cells have carbohydrate chains bound to themselves. Protein and lipid with carbohydrate chains are called glycoprotein and glycolipid, respectively. The carbohydrate chain structure has remarkable features different from nucleic acid and protein structures. A nucleic acid or a protein has a single binding pattern with a linear structure. In contrast, a carbohydrate chain has a ramified structure because the monosaccharide constituting the chain has several binding sites. Moreover, its component monosaccharides and chain length vary. Accordingly, the carbohydrate chain has an immensely complicated structure. The diversity of the carbohydrate chain structure is thought to have great significance for the carbohydrate chain function.

Figure 1:Carbohydrate chains on the surface of the cell membrane



Source: Partly compiled by the authors based on Reference[5]

2.2 Roles

The carbohydrate chain has three chief roles.

First, the carbohydrate chain helps its binding protein or lipid to obtain stability, adjust solubility, control intracellular localization and avoid decomposition by enzymes.

Second, the carbohydrate chain serves as an antenna in intercellular signaling. For instance, the carbohydrate chain plays a crucial role in intercellular recognition and cell adhesion. Also, viruses and bacteria recognize their targets of infection through carbohydrate chains of the hosts. In addition, the carbohydrate chain relates to the initiation of cell differentiation. Furthermore, the carbohydrate chain is involved in development, morphogenesis and fertilization, and, moreover, in proliferation and metastasis of cancer. Yet, most of these phenomena have not been analyzed at the molecular level.

Third, the carbohydrate chain helps in the recognition of individuals. Although the basic structure of the carbohydrate chain is preserved in each species, its additional structure varies among individuals or sometimes among populations. For example, the ABO blood types are determined by the difference of carbohydrate chains.

If we can analyze these carbohydrate chain functions at the molecular level, we can exploit the fruits of research in various fields such as the development of preventives against infections, control of infections and the immune system, application to cancer immunotherapy using carbohydrate chains, solution to immunological rejection in (artificial) organ transplantation and manipulation of development and differentiation.

3 Current status of carbohydrate chain research

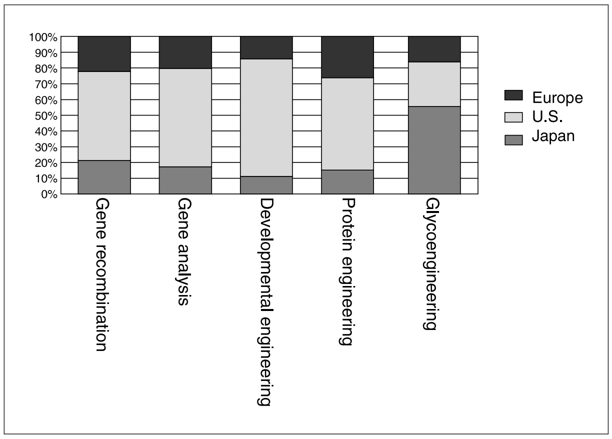
3.1 History of carbohydrate chain research[1-6]

In the late 1980s, it was clarified that the human erythropoietin, a regulator of erythrocyte production that is synthesized by bacteria through genetic engineering, is inactive because human-type carbohydrate chains are not bound to the erythropoietin produced by bacteria. Such research made the carbohydrate chain regarded as the third biological chain next to nucleic acid and protein.

Carbohydrate chain research is entering a new stage as cloning of the genes of carbohydrate-chain -related enzymes, i.e., enzymes for synthesizing, decomposing and modifying carbohydrate chains, has been successfully achieved one after another thanks to the development in the Human Genome Project and studies using these genes have become feasible. In other words, researchers have come to promote studies for unveiling the roles of the carbohydrate chain.

Japan has greatly contributed to the progress in carbohydrate chain research. Actually, the number of patent applications from Japan in glycoengineering is far larger than that in other fields of basic biotechnology and Japan accounts for 55% of all applications in glycoengineering (Figure 2). Meanwhile, it is thought that there are at least 300 genes related to glycosylation and researchers in the world are striving to clone these genes. As shown in Figure 3, 59% of the genes of glycosyltransferases have been cloned by Japanese research groups. In addition, Japan's advantage in carbohydrate chain research is supported by the high percentage of Japanese researchers making presentations in the two main international conferences related to the carbohydrate chain held once every two years. Japanese researchers accounted for 24% of all presenters in the International Carbohydrate Symposium held in Cairns, Australia, in 2002 and 28% in the International Symposium on Glycoconjugates held in The Hague, the Netherlands, in 2001.

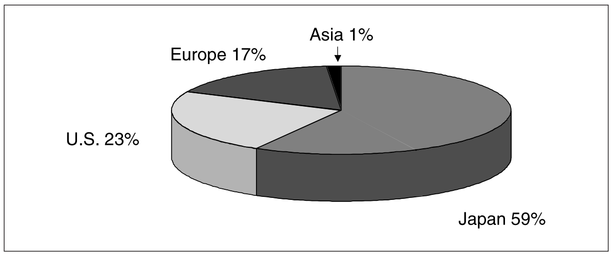
Figure 2:Comparison of nationality of patent applicants in basic biotechnology



Based on the analysis of patent applications filed all over the world with their priority dates between 1991 and 2000.

Source: Partly compiled by the authors based on Reference[7]

Figure 3: Percentage of the number of genes of glycosyltransferases cloned in each region or country



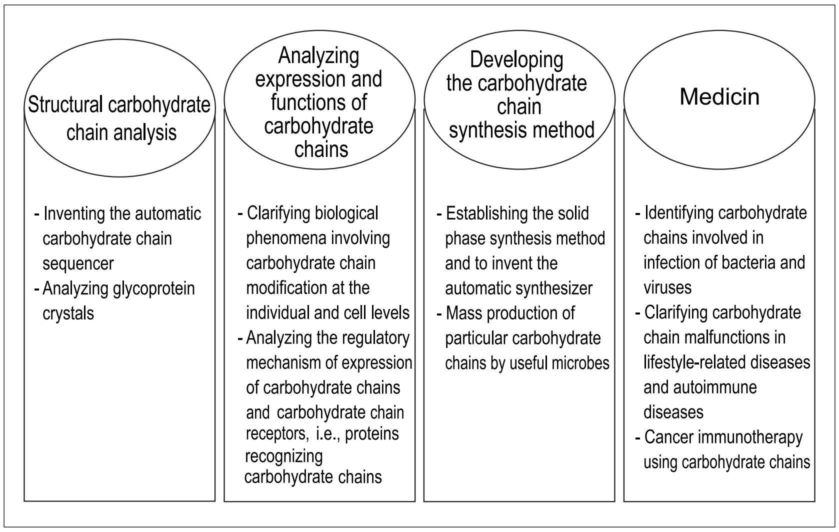
Source: Partly compiled by the authors based on information provided by Dr. Hisashi Narimatsu of the Research Center for Glycoscience, the National Institute of Advanced Industrial Science and Technology.

3.2 Current works

Carbohydrate chain research currently has four chief tasks as shown in Figure 4 and are tackled in the following projects.

The Japanese Ministry of Economy, Trade and Industry launched in FY2002 the "Glycoengineering

Figure 4:Research objectives in current projects


Structural carbohydrate chain analysis
- Inventing the automatic carbohydrate chain sequencer
- Analyzing glycoprotein crystals
Analyzing expression and functions of carbohydrate chains
- Clarifying biological phenomena involving carbohydrate chain modification at the individual and cell levels
- Analyzing the regulatory mechanism of expression of carbohydrate chains and carbohydrate chain receptors, i.e., proteins recognizing carbohydrate chains
Developing the carbohydrate chain synthesis method
- Establishing the solid phase synthesis method and to invent the automatic synthesizer
- Mass production of particular carbohydrate chains by useful microbes
Medicin
- Identifying carbohydrate chains involved in infection of bacteria and viruses
- Clarifying carbohydrate chain malfunctions in lifestyle-related diseases and autoimmune diseases
- Cancer immunotherapy using carbohydrate chains

Source: Partly compiled by the authors based onReference[1].

project (Development of structural carbohydrate chain analysis technology)" as a part of the "Biotechnology foundation research program for health maintenance and improvement". The ministry allocated about 1.1 billion yen for this project in the extra budget of FY2002 and 900 million yen in the FY2003 budget. This project aims to invent systems and devices for establishing novel structural carbohydrate chain analysis technology and the method for characterizing the carbohydrate chain structure, and for easily synthesizing standard samples of carbohydrate chains for carbohydrate chain characterization and glycoproteins necessary for functional analysis.

On the other hand, the Ministry of Education, Culture, Sports, Science and Technology has been advancing the "Support for promoting new industry using carbohydrate chain functions" as a part of the "R&D project for revitalizing the economy (Leading project)" and allocated one billion yen in the extra budget of FY2002. Also, the Core Research for Evolutional Science and Technology (CREST) of the Japan Science and Technology Corporation (JST) set the research area of "Clarification of carbohydrate chain structures and functions" and basic studies are promoted with research themes applied from the public from FY2002 to 2004.

Meanwhile, American and European researchers have also come to work vigorously on the research themes shown in Figure 4 along with an increasing interest in the carbohydrate chain function and the development of analyzers such as the mass spectrometer and fluorescence-activated cell sorter (FACS), and so on.

4 Tasks in carbohydrate chain research

In carbohydrate chain research, we need detailed molecular studies to identify functioning carbohydrate chains and to understand what substances the carbohydrate chains interact with and how the information is transmitted. Although the current carbohydrate chain studies have been improving as research projects proceed, very little analysis at the molecular level has been accomplished. If we do not have knowledge at the molecular level, we cannot effectively exploit carbohydrate chain functions. Indeed, why is the functional analysis of the carbohydrate chain at the molecular level so hindered? Glycolipid and glycoprotein researches have different hurdles. Yet, from a broader perspective, the difficulties in glycolipid studies coincide with those in glycoprotein studies. Therefore, we shall discuss the obstacles in glycoprotein studies in this chapter.

4.1 Necessity of an effective database of structure

Researchers have been trying to clarify what kind of carbohydrate chains a particular protein has by cutting out binding carbohydrate chains, isolating and purifying each carbohydrate chain and characterizing its structure. In characterization of the carbohydrate chain structure, an automatic sequencer of the carbohydrate chain has not been invented yet, enzymes cutting particular bindings need to be collected and the researcher needs to acquire complicated skills. Accordingly, in many cases, few fruits can be obtained in spite of the labor and expenses. Several months are often required to characterize the structure of one carbohydrate chain molecule. Since a glycoprotein usually has two to ten carbohydrate chains, it takes two to three years to characterize the structures of all carbohydrate chains of a glycoprotein. Data on about 50 carbohydrate chains of glycoproteins have been accumulated and part of them have been put into the database. Yet, the database lacks practicability because its scale is small and the search function for a carbohydrate chain with a particular structure is very limited.

4.2 Necessity of information on carbohydrate-chain binding positions

Upon characterizing carbohydrate chain structures, carbohydrate chains bound to a glycoprotein are cut out, isolated and purified. At that time, unfortunately, we cannot obtain positional information as to which amino acid of the protein each carbohydrate chain has been bound to, and, therefore, such data have been barely accumulated. For example, there are reports that the correct functioning of erythropoietin (a glycopeptide hormone) and the protein involved in immune deficiency syndrome depends on whether proper carbohydrate chains are bound to specific amino acids. The positional information as to where carbohydrate chains are bound is absolutely necessary in molecular studies on the carbohydrate chain function.

4.3 Necessity to identify functioning carbohydrate chains

In biological phenomena, the carbohydrate chain functions not only by itself but through cooperation with other proteins recognizing the information of the carbohydrate chain. Therefore, in functional carbohydrate chain analysis, we need to unveil the mechanism as to how the receptor molecule reads the information of the carbohydrate chain.

By the way, not every kind of carbohydrate chain functions; in fact, only a few carbohydrate chains function. It is often difficult to identify the carbohydrate chains that really function. Candidates of functioning carbohydrate chains sometimes cannot be selected from a group of carbohydrate chains that is complex and difficult to handle. In addition, several carbohydrate chains are sometimes related to one function, making it extremely difficult to identify the functioning carbohydrate chains.

If only the functioning carbohydrate chains can be identified, their mass production can become feasible and we can search for proteins by recognizing the information from them. Then, we can finally realize functional analysis of the carbohydrate chain at the molecular level to understand how information is transmitted between carbohydrate chains and proteins.

4.4 Summary of the hurdles in carbohydrate chain research

Structural characterization of each carbohydrate chain requires time and labor. Thus, if we can identify functioning carbohydrate chains and their binding positions in the protein prior to structural analysis, we can later conduct isolation, purification and structural characterization of carbohydrate chains that are not directly involved in functioning, thereby saving time and cost. In this way, we can efficiently characterize structures and functions of carbohydrate chains and, thereby, develop this field of study.

5 Necessity of research in glycoinformatics –A way to overcome the hurdles in carbohydrate chain research

If we can clear the obstacles as described in Chapter 4, that is, if we can establish the methodology for identifying carbohydrate chain binding positions and functioning carbohydrate chains, molecular analysis of carbohydrate chain functions will advance and carbohydrate chain research will be boosted. For this purpose, we need to approach glycoinformatics, a new field of study.

5.1 Concept of glycoinformatics

Glycoinformatics aims to promote structural and functional carbohydrate chain research and its applied studies by obtaining, accumulating and exploiting data on both structures and functions of carbohydrate chains. The basic technology of bioinformatics can be applied to this field of study.

Bioinformatics is defined as research, development and application of computer tools and approaches for obtaining, accumulating, systematizing and analyzing data on biology, medicine, ethology and health, making their database and developing the data including their visualization.[8] Today, bioinformatics mainly deals with DNA and protein, and researchers are compiling databases and developing forecasting programs. Bioinformatics is advancing other fields of life science as well.

Although the term glycoinformatics has not yet become popular, some researchers are endeavoring to establish databases to bolster up glycoinformatics. This move can be seen not only in Japan. As we see the resumes of meetings of societies concerning mass spectroscopy, we can surmise that the research group led by H. Freeze of The Burnham Institute, for example, has begun studies with an eye to the compilation of a new database of carbohydrate chains.

5.2 Technical tasks in developing glycoinformatics

We need to establish the methodology for identifying carbohydrate-chain binding positions and functioning carbohydrate chains and compile a database of carbohydrate chains in order to overcome the hurdles in carbohydrate chain research. It is necessary to compile a database with which the glycome (i.e., the total of carbohydrate chains of the cell and individual) can be analyzed and compared between cells or individuals. Although a database including detailed carbohydrate chain structures would be useful, its compilation will require much time due to the difficulty in structural carbohydrate chain analysis. Therefore, it is our technical task to determine what kind of information to collect.

Now we have two tasks: 1. Determining what kind of structural information to accumulate; 2. Considering and developing a way to put data into the database and search for information from the database, and studying how to operate the database.

In the first task, the conventional method to characterize each carbohydrate chain structure requires much time and money as described above, so it is impractical to collect data on carbohydrate chain structures themselves. Thus, we need to scrutinize as to what we should use as information substituting for the carbohydrate chain structure that can be obtained speedily with small quantities of samples.

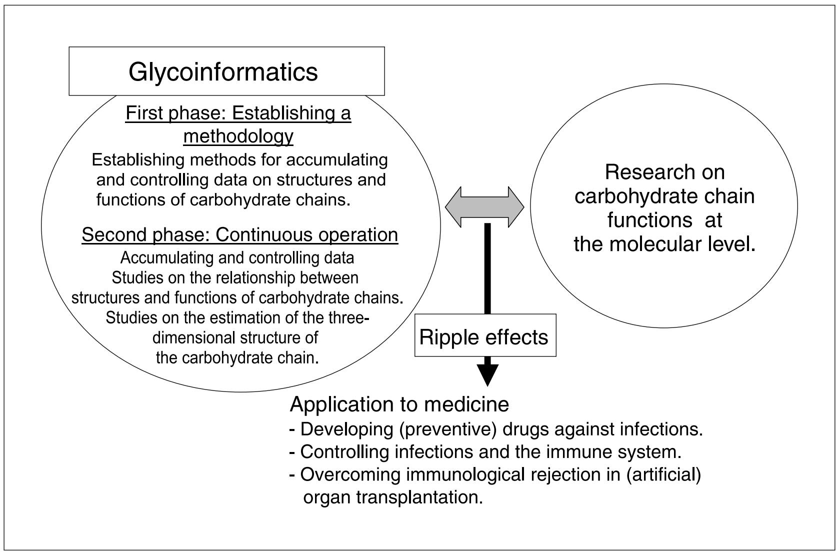
For instance, we may use mass spectrometric data of glycoproteins. We can obtain information on the carbohydrate chain binding position in the protein, molecular weight of the carbohydrate chain and component monosaccharide sequences by using the matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometer*1, which is thought to be useful in glycoprotein analysis. However, isomers of monosaccharides cannot be distinguished by the mass spectrometer, so we need additional information to recognize them. For example, we may use several proteins that distinguish isomeric structures of carbohydrate chains and bind to them at different strength depending on their structures. Lectin*2, antibodies against carbohydrate chains and enzymes synthesizing or cutting carbohydrate chains are examples of such proteins. We can distinguish between isomers by comparing the affinity between each protein and carbohydrate chain.

On the other hand, the problem in accomplishing the second task above is not the technical issues but is the fact that it has not been dealt with until now. This task needs to be tackled in cooperation with information scientists.

5.3 Benefit of glycoinformatics

When glycoinformatics has proceeded and a basic database has become solid, structural and functional information on carbohydrate chains can be applied to the full extent. Then, functional carbohydrate chain analysis at the molecular level will progress and existent

Figure 5:Glycoinformatics and its roles


Glycoinformatics
First phase: Establishing a methodology
Establishing methods for accumulating and controlling data on structures and functions of carbohydrate chains.
Second phase: Continuous operation
Accumulating and controlling data
Studies on the relationship between structures and functions of carbohydrate chains.
Studies on the estimation of the three-dimensional structure of the carbohydrate chain.
Research on carbohydrate chain functions at the molecular level.
Ripple effects
Application to medicine
- Developing (preventive) drugs against infections.
- Controlling infections and the immune system.
- Overcoming immunological rejection in (artificial) organ transplantation.

research projects as shown in Figure 4 will be boosted further.

As a consequence, it will become possible to develop (preventive) drugs against infections, control infections and the immune system, improve cancer immunotherapy using carbohydrate chains, overcome immunological rejection in (artificial) organ transplantation, manipulate development and differentiation and develop artificial receptors of carbohydrate chains. Also, if it becomes possible to accumulate information on glycomes, or the total of carbohydrate chains, of individuals and trace them temporally, a new type of individual health screening will be realized. Carbohydrate chains change dynamically in the progress of cancer, aging, development and differentiation. Accordingly, by temporally tracing the glycome data of individuals, we can realize a new monitoring method for detecting cancers, clarifying metastasis of cancer or estimating the progress of aging.

5.4 Projects for developing glycoinformatics

We can divide the process for the practical foundation and development of glycoinformatics into two phases as follows.

(1) First phase: establishing the methodology

In this phase, we modify the mass spectrometer for obtaining structural information of the carbohydrate chain and determine what kind of data to collect and how to accumulate them in the database. The final goal is to establish a database where data on both structures and functions of carbohydrate chains can be searched and used.

To achieve this objective, researchers of not only glycoscience but also other fields must attack the problems together and seek effective measures. We need to draw up an intensive project plan preferably for five to eight years with 40 to 50 researchers participating.

(2) Second phase: continuous operation

Data are accumulated in the database, controlled and utilized in this phase. As the DNA Data Bank of Japan (DDBJ) accumulates and controls the data on DNA sequences, we need to consider establishing and administering a database management organization that accumulates and controls the data on the structural and functional information on carbohydrate chains.

6 Conclusion

As the post-genome era has opened, carbohydrate chain research is entering a new stage toward functional analysis at the molecular level. In such a situation, we need to promote glycoinformatics, a new field of study. Glycoinformatics will accelerate functional carbohydrate chain analysis at the molecular level.

The important thing is to handle not just a part of carbohydrate chains but carbohydrate chains as a whole, i.e., glycome of the cell and individual. We need to obtain comprehensive data rather than only detailed data. Informatics and the database with which such data can be compared and studied will help molecular analysis.

In carbohydrate chain research, Japan has high potential and has been a front-runner in the world. In order to make good use of such potential, we must promote glycoinformatics to boost carbohydrate chain research even further.

Glossary

*1 Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometer

This ionization method was invented by Koichi Tanaka of Shimazu Corporation and enabled mass spectroscopy of biological polymers. Development of this technology contributed to Tanaka's winning of the Nobel Prize in chemistry in 2002.

*2 Lectin

Herman Stillmark discovered this protein in 1888 from the fact that extract from castor beans serve as agglutinin of erythrocytes of various animals. Lectin is the total of proteins that recognizes specific structures of carbohydrate chains.

References

[1] "Sugar chains as the Third Biomolecule, and Post-genome Researches," Science & Technology Trends – Quarterly Review No. 4., December 2002.

[2] "Carbohydrate chain: mystery of life beyond genomic information." Memoir of the 16th symposium of "University Science," Kuba Pro., 2003. (in Japanese)

[3] "Functions of Glyco-chains: As the Third Chain Molecule Next to Nucleic Acid and Protein," Protein, Nucleic Acid and Enzyme (extra number), 2003. (in Japanese)

[4] "Essentials of Glycobiology," edited by Varki, A. et al., Cold Spring Harbor Laboratory Press, 1999.

[5] "Feature article: How does the carbohydrate chain determine the biological function?," edited by Tsuji, S., Cell Technology, Vol. 15, No. 6, 1996. (in Japanese)

[6] "Handbook of Glycosyltransferases and Related Genes," edited by Taniguchi, N. et al., Springer-Verlag, Tokyo, 2002.

[7] "Trends of Industrial Property Right Applications and Registrations in Life Science," Technological Trend Research Section, Technology Research Division, General Affairs Department, Japan Patent Office, April 24, 2003. (in Japanese)

[8] "Bioinformatics," edited by Sugawara, H., Kyoritsu Shuppan Co., Ltd., 2002. (in Japanese)

(Original Japanese version: published in September 2003)