Abstract: One hundred ten genes for human glycosyltransferases had been cloned and analyzed at the beginning of April, 2001 after the first mammalian glycosyltransferase gene was cloned in 1986. The term glycogene includes the genes for glycosyltransferases, sulfotransferases adding sulfate to carbohydrates and sugar-nucleotide transporters, etc. In April 2001, we started the Glycogene Project (GG project), which was a comprehensive study on human glycogenes. One hundred five novel glycogenes were identified as candidates with the aid of bioinformatic technology. All of them were cloned and expressed as recombinant enzymes, and their substrate specificities were then examined using various acceptors. Thirty-eight glycogenes among the 105 candidates were determined to be glycosyltransferases, sulfotransferases and sugar-nucleotide transporters. One hundred sixty-five glycogenes were subcloned into a Gateway entry vector, and prepared as a human glycogene library. These cloned glycogenes can be easily expressed as recombinant enzymes in various expression systems.

Key words: glycogene, glycogene project, glycosyltransferase, sulfotransferase, sugar-nucleotide transporter

Bioinformatics was powerful to identify novel glyco- genes in the databases.

We constructed a system for predicting novel glycosyltransferase genes from the primary sequence characteristics shared by glycosyltransferases. This system has functions such as an automated BLAST search, automatic removal of known sequences, assembly of EST search results (EST sequence hits are assembled with Phrap to eliminate duplicate hits for better search efficiency), and prediction of gene regions in genome sequence search results (in the genome sequence search, gene regions are predicted with GENSCAN, and the amino acid sequence for the entire ORF is obtained), in combination with the program described below to identify glycosyltransferases. The general characteristics of glycosyltransferases are as follows: 1) The N-terminus is short and within the cytoplasm; 2) The trans-Golgi membrane region (a sequence of approximately 18–20 residues rich in hydrophobic amino acids) is in the vicinity of the amino terminus; 3) This region is followed by the stem domain, which is expected to dramatically advance:

1) the elucidation of the functions of sugar chains as bioactive substances including carrier molecules, 2) the elucidation of control mechanisms of carrier molecules by sugar chains, 3) the development of analytical techniques for sugar chain structures, and 4) the development of automated synthesis techniques for sugar chains.

The project Construction of a Human Glycogene Library and Comprehensive Functional Analysis (the GG project) was supported by the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry (METI) of Japan, and performed by us from April, 2001, to March, 2004.

A new term ‘glycogene’ was given to the genes involved in glycosylation of proteins, lipids and proteoglycans. Glycogenes include the genes for 1) glycosyltransferases, 2) glycolytic enzymes, 3) sugar nucleotide synthetases, 4) sugar nucleotide transporters, and, in a broader sense, sugar chain-recognizing molecules such as 5) lectins.

Human genome analysis has almost been completed, and the number of the genes was estimated to be about 30,000; that is quite a bit smaller than the predicted number. We propose that posttranslational modifications such as phosphorylation and glycosylation are important to increase the protein function. Suppose that different structures of glycosylation are on a single protein; glycosylation may confer different functions to the protein. Thus, the consideration of protein functions should include posttranslational modifications.

One gene is involved in the synthesis of the protein moiety of a glycoprotein, whereas dozens of genes are involved in the synthesis of its sugar chain moiety. Thus, a single glycoprotein is the joint product of dozens of genes. We speculate that human beings could evolve from lower eukaryotes with a small increase in the number of genes; however, each protein function of higher eukaryotes might be more intricately regulated by glycosylation than that of lower eukaryotes.

To understand the complex structures and functions of glycosylation, we need to obtain all glycogenes on hand, because one structure of glycosylation is formed by cooperative and stepwise reaction of many glycogenes. Therefore, we started the comprehensive identification and functional analysis of human glycogenes. These are expected to dramatically advance:

1) the elucidation of the functions of sugar chains as bioactive substances including carrier molecules, 2) the elucidation of control mechanisms of carrier molecules by sugar chains, 3) the development of analytical techniques for sugar chain structures, and 4) the development of automated synthesis techniques for sugar chains.

The project Construction of a Human Glycogene Library and Comprehensive Functional Analysis (the GG project) was supported by the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry (METI) of Japan, and performed by us from April, 2001, to March, 2004.

Bioinformatics was powerful to identify novel glycogenes in the databases.

We constructed a system for predicting novel glycosyltransferase genes from the primary sequence characteristics shared by glycosyltransferases. This system has functions such as an automated BLAST search, automatic removal of known sequences, assembly of EST search results (EST sequence hits are assembled with Phrap to eliminate duplicate hits for better search efficiency), and prediction of gene regions in genome sequence search results (in the genome sequence search, gene regions are predicted with GENSCAN, and the amino acid sequence for the entire ORF is obtained), in combination with the program described below to identify glycosyltransferases. The general characteristics of glycosyltransferases are as follows:

1) The N-terminus is short and within the cytoplasm; 2) The trans-Golgi membrane region (a sequence of approximately 18–20 residues rich in hydrophobic amino acids) is in the vicinity of the amino terminus; 3) This region is followed by the stem domain, which is expected to dramatically advance:

1) the elucidation of the functions of sugar chains as bioactive substances including carrier molecules, 2) the elucidation of control mechanisms of carrier molecules by sugar chains, 3) the development of analytical techniques for sugar chain structures, and 4) the development of automated synthesis techniques for sugar chains.

The project Construction of a Human Glycogene Library and Comprehensive Functional Analysis (the GG project) was supported by the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry (METI) of Japan, and performed by us from April, 2001, to March, 2004.

Bioinformatics was powerful to identify novel glycogenes in the databases.

We constructed a system for predicting novel glycosyltransferase genes from the primary sequence characteristics shared by glycosyltransferases. This system has functions such as an automated BLAST search, automatic removal of known sequences, assembly of EST search results (EST sequence hits are assembled with Phrap to eliminate duplicate hits for better search efficiency), and prediction of gene regions in genome sequence search results (in the genome sequence search, gene regions are predicted with GENSCAN, and the amino acid sequence for the entire ORF is obtained), in combination with the program described below to identify glycosyltransferases. The general characteristics of glycosyltransferases are as follows:

1) The N-terminus is short and within the cytoplasm; 2) The trans-Golgi membrane region (a sequence of approximately 18–20 residues rich in hydrophobic amino acids) is in the vicinity of the amino terminus; 3) This region is followed by the stem domain, which is expected to dramatically advance:

1) the elucidation of the functions of sugar chains as bioactive substances including carrier molecules, 2) the elucidation of control mechanisms of carrier molecules by sugar chains, 3) the development of analytical techniques for sugar chain structures, and 4) the development of automated synthesis techniques for sugar chains.

The project Construction of a Human Glycogene Library and Comprehensive Functional Analysis (the GG project) was supported by the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry (METI) of Japan, and performed by us from April, 2001, to March, 2004.

Bioinformatics was powerful to identify novel glycogenes in the databases.

We constructed a system for predicting novel glycosyltransferase genes from the primary sequence characteristics shared by glycosyltransferases. This system has functions such as an automated BLAST search, automatic removal of known sequences, assembly of EST search results (EST sequence hits are assembled with Phrap to eliminate duplicate hits for better search efficiency), and prediction of gene regions in genome sequence search results (in the genome sequence search, gene regions are predicted with GENSCAN, and the amino acid sequence for the entire ORF is obtained), in combination with the program described below to identify glycosyltransferases. The general characteristics of glycosyltransferases are as follows:

1) The N-terminus is short and within the cytoplasm; 2) The trans-Golgi membrane region (a sequence of approximately 18–20 residues rich in hydrophobic amino acids) is in the vicinity of the amino terminus; 3) This region is followed by the stem domain, which is expected to dramatically advance:

1) the elucidation of the functions of sugar chains as bioactive substances including carrier molecules, 2) the elucidation of control mechanisms of carrier molecules by sugar chains, 3) the development of analytical techniques for sugar chain structures, and 4) the development of automated synthesis techniques for sugar chains.

The project Construction of a Human Glycogene Library and Comprehensive Functional Analysis (the GG project) was supported by the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry (METI) of Japan, and performed by us from April, 2001, to March, 2004.
rich in proline (an amino acid forming a non high-order structure) and in the O-glycan-linked amino acid residues serine and threonine; 4) The C-terminus is followed by an approximately 300-400-residue catalytic site; 5) The catalytic site contains DxD or DxH, an amino acid sequence necessary for binding divalent cations. In addition to these characteristics, several motif regions were set up for every existing gene family using MEME, then a program was designed to assess six items including whether the motif regions are present, and has been used to determine possible glycosyltransferases. However, since this system is only capable of detecting genes with some degree of homology to the existing ones, we are in the process of expanding its functions, such as the incorporation of software capable of comparing high-order structures. These systems are sufficiently applicable to the analysis of not only glycosyltransferase genes but also various other genes including lectin genes.

**Current status of glycogene research.**

Since the first mammalian glycosyltransferase gene was cloned by two groups independently, more than 160 human glycosyltransferase genes have been cloned to date. Figure 1 summarizes the number of human glycogenes which were reported to date. In total, 166 glycogenes were reported, mostly by Japanese researchers. The numbers in parentheses indicate the glycogenes which were analyzed and reported in the GG project. Sixty-eight candidate genes still remain to be analyzed for their activity. Nineteen genes were found to be pseudogenes. Thirty-eight glycogenes in total were reported by the GG project. The details of each gene will be referenced in the original papers published by us. The sequence of studies is as follows: 1) Glycogenes are cloned and the encoded enzymes are expressed by various methods; 2) The enzymes are analyzed for the synthesis of sugar chain structures in vitro or in vivo. Subsequently, the synthesized sugar chain structures are examined for their biological functions as follows: 3) Glycogenes are transfected into cultured cells, and the functions of the cells whose sugar chain structures have undergone changes are analyzed.

The characteristics of a few glycosyltransferase gene families and their substrate specificities are presented below along with the results of the GG project. For more details, the reader is referred to our cited publications.

**Glycosyltransferase gene family with β4-glycosyltransferase (β4GT) motif.**

The β4-galactosyltransferase (β4Gal-T) gene was the first glycosyltransferase cloned. β4Gal-T forms a family including seven enzymes, β4Gal-T1 through β4Gal-T7, which share the same characteristic of transferring galactose from the sugar donor UDP-Gal to the sugar acceptor via a β1,4 linkage, but differ in using sugar acceptors such as sugar chains in glycoproteins, glycolipids, or glycosaminoglycans. Using the amino acid sequences of this family as query sequences, we performed database searches and found eight new members sharing a β4GT motif, GWGXED. Six of them were finally determined to be the genes involved in the synthesis of chondroitin sulfate. Chondroitin sulfate occurs as a sugar chain of proteoglycans in which dozens of the disaccharide repeating unit (GlcA β1,3GalNAc β1,4-) are bound to a serine residue of the core protein via a linkage tetrasaccharide (GlcA β1,3,Gal β1,3,Gal β1,4Xyl). In addition, two novel members were found recently and their substrate specificity was analyzed. They were identified as the enzymes responsible for the synthesis of the N-acetylatedidactosamine (LacdiNAc), GalNAc β1,4GlcNAc, structure. The LacdiNAc structure on N-glycans was found on some specific glycoproteins which are glycoprotein hormones. The LacdiNAc structure was reported to determine the half-life time of these hormones in blood.

All members in the β4GT family, the β4-galactosyltransferase gene, the chondroitin sulfate-synthesizing enzyme gene and the LacdiNAc synthase families, have the amino acid sequence GWGXED in common, but differ in their sugar donors and acceptors, suggesting that this motif is the sequence specifying the β1,4 linkage.

**Glycosyltransferase gene family with β3-glycosyltransferase (β3GT) motif.**

The glycosyltransferase gene family with the β3 motif includes five β3-galactosyltransferases (β3Gal-T), six β3-N-acetylgalcosaminyltransferases (β3Gn-T) and two β3-N-acetylgalactosaminyltransferases (β3GalNAc-T).

Of the many β3Gn-Ts reported by us, β3Gn-T2 clearly shows the strongest activity for the sugar acceptor polyaccharide structures, suggesting that this is a polyaccharide-synthesizing enzyme on glycoproteins. We performed an in silico analysis using the database, successfully cloned the Lac-Cer-synthesizing enzyme gene, β3Gn-T3, and the O-liked sugar-chain core β3-synthesizing enzyme gene, β3Gn-T6, and reported their substrate specificities. β3Gn-T6 is restrictively localized in the epithelia of the stomach and colon, in which the major form of O-glycans is the core 3 structure. This enzyme may be profoundly involved in cancer metastasis. A novel enzyme, named β3GalNAc-T2, was found by in silico cloning and characterized. β3GalNAc-T2 transfers GalNAc to GlcNAc with a β3-linkage on the termini of N- and O-glycans. Although the GalNAc β1,3 GlcNAc β1-R structure has not been reported in humans
or other mammals, it must exist where the enzyme is expressed. Very recently, β3Gn-T8 was identified as being responsible for the synthesis of polylactosamine chains on β1-6 branched N-glycans, and dramatically up-regulated in cancer tissue.22)

**Polypeptide N-acetylgalactosaminyltransferase gene family.**

Polypeptide N-acetylgalactosaminyltransferase (pp-GalNAc-T) is a group of glycosyltransferases transferring GalNAc to serine or threonine residues in O-glycans through an α-linkage. At the present time, 15 members of this family have been published. We cloned five of these genes, pp-GalNAc-T10, -T12, -T13, -T14 and -T15, and analyzed their functions.23-26) The sequences of this enzyme family are characterized by the presence of conserved motifs called the GT1 and Gal/GalNAc-T motifs, and have homology throughout the entire lengths. Thus, it is easy to find homologous members in the databases. Besides the 15 members published, an additional five members can be found in the databases. However, the activity of these five candidates has not been detected yet.

Each pp-GalNAc-T exhibits different substrate specificity for the peptide sequence, and shows different tissue distribution. Thus, initiation of O-glycosylation which is determined by the activity of pp-GalNAc-T is quite complicated in vivo. An example of such analyses is the case of multiple O-linked sugar chains in the hinge region of IgA-1. Based on the tissue distribution of expression of the pp-GalNAc-T family and the differences in their relative activity determined with the hinge region peptide, we reported that pp-GalNAc-T2 transfers N-acetylgalactosamine to the IgA-1 hinge region.27)

**Conclusion.**

At present, sugar-chain engineering has been given the status of one of the key techniques of biotechnology in Japan. In fact, Japan is most advanced in the glycogene discovery for these years as shown in Fig. 2. At the beginning of the GG project, Japan accounted for 50% of the glycogenes. The percentage increased to 61% for these three years because of the success of the GG project.

**REFERENCES**

14) T. Sato, M. Gotoh, K. Kiyohara, A. Kameyama, T. Kubota, N.


S. Kamiyama, T. Suda, M. Suzuki, N. Kiakaya, H. Nishihara: Molecular cloning and characterization of a human multido- Static content...
ポストゲノム・プロジェクト：
ヒト糖鎖遺伝子の網羅的解析

成松 久
産総研総合研究所糖鎖工学研究センター
糖鎖遺伝子機能解析チーム
(305-8568 つくば市栄園 1-1-1 中央第2 OSL-C2)
1986年に最初の哺乳類の糖転移酵素遺伝子がクローニングされて以来、2001年4月現在までに、110種類のヒト糖転移酵素遺伝子がクローニングされ解析されていった。糖鎖遺伝子という語句は、糖転移酵素、糖に硫酸基を転移する硫酸転移酵素、糖スクレオチドトランスポーターなどの遺伝子を含む。2001年4月に、我々は、ヒト糖鎖遺伝子の網羅的解析（糖鎖遺伝子プロジェクト；GGプロジェクト）を開始した。バイオインフォマティクスの技術を活用して、105種類の新規な候補遺伝子を同定した。すべてをクローニングし、リコンピナント酵素として発現し、それらの特異性をあらゆるアクセプター基質を用いて解析した。105候補のうち、38糖鎖遺伝子は糖転移酵素、硫酸転移酵素、糖スクレオチドトランスポーターであることを決定した。165の糖鎖遺伝子を、ゲートウェイのエンタープライズクローンンシステムを用いてクローンし、糖鎖遺伝子ライブラリとして整備した。このクローン化された糖鎖遺伝子は、それ以外で簡便に様々な発現系を用いて、リコンピナント酵素として発現できる。