An introduction to bioinformatics for glycomics research

Kiyoko F. Aoki-Kinoshita, Ph.D.
Dept. of Bioinformatics
Faculty of Engineering, Soka University
kkiyoko@soka.ac.jp

Abstract:

Carbohydrates are the third class of information-encoding biological macromolecules. The term 'glycomics,' the scientific attempt of characterizing and studying carbohydrates, is a rapidly emerging branch of science, for which informatics is just beginning. Glycomics requires sophisticated algorithmic approaches. The status of structural encoding, standardization, databases and applications will be presented.

Goals

- To introduce the role of carbohydrates as the third class of information-containing biological macromolecules to a broader audience of computational biologists.
- To explain the fundamental characteristics of carbohydrate structures, which require specific encoding schemata as well as algorithmic developments.
- To give an overview of available database resources, algorithms and techniques.
- To outline a systems approach to discover the structure-function relationships of glycans.

Schedule:

- General Introduction: (60 minutes plus 15 minutes for discussion)
  - Biological role of carbohydrates as information containing molecules (15 minutes)
  - The role of informatics in glycomics research (15 minutes)
  - The challenges to connect glyco-related databases (30 minutes)
- Glycome Informatics: (120 minutes plus 15 minutes for discussion)
  - Computer theoretic algorithms for tree structures (30 minutes)
  - Automated MS prediction algorithms (20 minutes)
  - Data mining of glycan structures using kernels and probabilistic models (40 minutes)
  - Systems approaches to unveiling structure-function relationships of glycans (30 minutes)

Overview:

- General Introduction: (60 minutes plus 15 minutes for discussion)
  - Biological role of carbohydrates as information containing molecules (30 minutes)

    This section will briefly summarize the biological pathways by which carbohydrates are synthesized, their structural variability found in nature in various species as well the biological functions in which they are involved. The diseases in which carbohydrates are given as potential drug targets will be summarized.
  - The role of informatics in glycomics research (15 minutes)

    This section introduce the challenges of glycomics research for experimentalists as well as computer scientists; summarize the current status of achieved standardization as well as already available bioinformatics applications in glycobiology. It will outline the most urgent next steps which are fundamental for rapid progress in this area.
  - The challenges to connect glyco-related databases (15 minutes)

    No standard databases exist where all published glycan structures found in various species, organs, tissues or cells can be routinely retrieved. A history of the development of glyco-related databases will be presented. The existing retrieval concepts of the publicly available (GLYCOSCIENCES.de, KEGG Glycan, CFG, EuroCarbDB, GlycomeDB) databases will be discussed. The needs for standardization especially for the exchange of glycan structures will be discussed. An outline will be given about the expected synergetic effects when establishing an efficient connection of the existing resources.
- Glycome Informatics: (120 minutes plus 15 minutes for discussion)
  - Computer theoretic algorithms for tree structures (30 minutes)

    Computer theoretic methods applied to glycans will be introduced in this section. Glycan structures can most readily be represented by ordered labeled trees, where the labels of the nodes are monosaccharides and the labels of the edges are the glycosidic bond conformations. As such, several algorithms for studying glycans have recently been developed, including dynamic programming methods for tree structure alignment and resulting score matrices, which have produced very interesting and promising results.
  - Automated MS prediction and probabilistic models (20 minutes)

    The foremost challenge in glycomics is the efficient annotation of glycan structures from mass spectral data. Several methods will be introduced that theoretically predict glycan structures from MS data or vice versa.
  - Data mining of glycan structures using probabilistic models and kernels (40 minutes)

    Advanced data mining methods for labeled ordered trees, and in particular for capturing "sibling-dependent" patterns across the breadth of the leaves, have been developed. This latter model has shown promise to capture the patterns of glycan structures being recognized by lectins. Additionally, kernel methods are also proving to be useful for extracting cell-specific glycan structures, with the potential to predict glycan markers for specific diseases.

- Systems approach to unveiling structure-function relationships of glycans (30 minutes)

  Glycans are more diverse in terms of chemical structure and information density than are DNA and proteins. This diversity arises from glycans' complex non-template based biosynthesis, which involves several enzymes and isoforms of these enzymes. Thus, glycans are expressed as an 'ensemble' of structures that mediate function. Moreover, unlike protein-protein interactions, which can be generally viewed as 'digital' in regulating function, glycan-protein interactions impinge on biological functions in a more 'analog' fashion that can in turn 'fine-tune' a biological response. This fine-tuning by glycans is achieved through the graded affinity, avidity and multivalency of their interactions. This section focuses on areas of technologies and the importance of developing a bioinformatics platform to integrate the diverse datasets generated using the different technologies to allow a systems approach to glycan structure-function relationships.

Mammalian Glycan Linkages Produced by Glycosylation

There are nine nucleotide sugar donors and multiple protein and lipid acceptor motifs for glycosyltransferases, which produce 14 different glycans in stereoisomeric configurations (α or ß) linked at the number 1 position of the donor sugar ring. The attached mono-saccharide frequently then becomes a saccharide acceptor in 1 of 49 other glycosyltransferase reactions. This results in glycosidic bonds with α or ß configurations of the donor saccharide linked through position 1 or 2 to position 2, 3, 4, or 6 of an acceptor saccharide. Glycan diversification is dictated by the combinatorial and regulated application of this enzymatic potential.This includes hyaluronan synthesis, which occurs by copolymerization of two nucleotide sugar donors. The formation of iduronic acid by the epimerization of glucuronic acid subsequent to glycosylation is not depicted. Although some disaccharide sequences are found on multiple types of glycans, others are specific to one or few glycan types.

Cellular Regulation of Glycan Expression

Multiple mechanisms that alter cellular glycosyltransferase or glycosidase expression, structure, and activity, which can thereby regulate the formation of glycans, are represented. These include (1) control of glycosyltransferase and glycosidase gene transcription, (2) synthesis and transport of nucleotide sugar donors to the ER and Golgi (sugar transporters not depicted), (3) modulation of enzymatic structure through phosphorylation, (4) relative amounts of enzymes that compete for identical substrates, (5) intra-cellular enzyme trafficking and altered access to substrates, (6) proteolysis within the lumen of the Golgi resulting in secretion of catalytic domains, and (7) glycan turnover at the cell surface by endocytosis coincident with expression of different glycans from altered glycan synthesis. Effects of glycosyltransferase and glycosidase cytoplasmic tail phosphorylation (3) and intraluminal proteolysis (6) on cellular glycosylation remain to be established.

Overview of glycan biosynthetic pathways

The **N-linked pathway** begins in the endoplasmic reticulum (ER) with transfer of a preformed glycan to glycosylation sites soon after the nascent protein emerges from the ribosome into the ER lumen. Glycan remodeling begins in the ER, but mostly occurs in the Golgi.

The first step of **O-mannose-linked** glycan assembly occurs in the ER and is then completed in the Golgi, whereas all of the O-xylose and O-GalNAc (O-linked N-acetylgalactosamine) pathway steps occur in the Golgi.

GPI (glycophospholipid) anchors are preassembled in the ER. They are added near the C termini of specific proteins soon after their synthesis.

The synthesis of **glycosphingolipids (GSLs)** begins on the cytoplasmic side of the ER, after which the nascent glycan flips into the lumen and is extended with additional sugars in the Golgi. After leaving the Golgi, glycan-decorated proteins and lipids traffic to the cell surface, lysosome or other vesicles; trafficking to the plasma membrane only is shown here for simplicity. GPI-linked proteins and GSLs both cluster in similar regions of the plasma membrane as both are associated with lipid rafts.

A substantial number of glycan-carrying proteins are subsequently secreted into the extracellular matrix.

N-Linked Glycan Biosynthesis

1. Monosaccharides of GlcNAc (N-acetylglucosamine) and mannose are
   a. activated or interconverted, and
   b. used to build the lipid-linked oligosaccharide precursor (LLO), which is assembled on dolichol (a polyisoprenoid derived from the cholesterol pathway)
2. Dolichol in the endoplasmic reticulum (ER) membrane is activated to dolichol phosphate by a kinase.
3. UDP-GlcNAc provides GlcNAc-1-phosphate and GlcNAc,
4. GDP-mannose provides 5 mannose residues that are added on the cytoplasmic side of the ER.
5. The resulting mannose5-GlcNAc2-P-P-dolichol flips into the ER lumen
6. 4 mannose and 3 glucose residues are added using dolichol-P-mannose and dolichol-P-glucose donors, respectively.
7. An oligosaccharyltransferase complex delivers the glycan to asparagine residues in the nascent protein core.
8. The protein-bound glycan is trimmed to mannose8-GlcNAc2-protein in the ER and is trimmed further in the Golgi.
9. One GlcNAc residue is added to the trimmed mannose5 structure
   a. This GlcNAc can be extended with galactose and sialic acid to form a hybrid glycan or,
   b. by adding more GlcNAc, branching can continue to form complex-type oligosaccharides.
10. Nucleotide sugar transporters in the Golgi provide substrates for glycan extension.

Human diseases caused by genetic defects in N-glycosylation pathways

Disorder	Gene	Enzyme	OMIM	Key Features
CDG-Ia	PMM2	Phosphomannomutase II	212065	Mental retardation, hypotonia, esotropia, lipodystrophy, cerebellar hypoplasia, stroke-like episodes, seizures
CDG-Ib	MPI	Phosphomannose isomerase	602579	Hepatic fibrosis, protein-losing enteropathy, coagulopathy, hypoglycaemia
CDG-Ic	ALG6	Glucosyltransferase I Dol-PGlc: Man_9-$GlcNAc_2$-P-P-Dol glucosyltransferase	603147	Moderate mental retardation, hypotonia, esotropia, epilepsy
CDG-Id	ALG3	Dol-P-Man:Man_5-$GlcNAc_2$-P-P-Dol mannosyltransferase	601110	Profound psychomotor delay, optic atrophy, acquired microcephaly, iris colobomas, hypsarrhythmia
CDG-Ie	DPM1	Dol-P-Man synthase I GDP-Man: Dol-P-mannosyltransferase	603503	Severe mental retardation, epilepsy, hypotonia, mild dysmorphism, coagulopathy
CDG-If	MPDU1	Man-P-Dol utilization 1/Lec35	608799	Short stature, icthyosis, psychomotor retardation, pigmentary retinopathy
CDG-Ig	ALG12	Dol-P-Man:Man_7-$GlcNAc_2$P-P-Dol mannosyltransferase	607143	Hypotonia, facial dysmorphism, psychomotor retardation, acquired microcephaly, frequent infections
CDG-Ih	ALG8	Glucosyltransferase II Dol-P-Glc: Glc_1-Man_9-$GlcNAc_2$-P-P-Dol glucosyltransferase	608104	Hepatomegaly, protein-losing enteropathy, renal failure, hypoalbuminaemia, oedema, ascites
CDG-Ii	ALG2	Mannosyltransferase II GDPMan: Man_1-$GlcNAc_2$-P-P-Dol mannosyltransferase	607906	Normal at birth; mental retardation, hypomyelination, intractable seizures, iris colobomas, hepatomegaly, coagulopathy
CDG-Ij	DPAGT1	UDP-GlcNAc: Dol-P-GlcNAc-P transferase	608093	Severe mental retardation, hypotonia, seizures, microcephaly, exotropia
CDG-Ik	ALG1	Mannosyltransferase I GDPMan: $GlcNAc_2$-P-P-Dol mannosyltransferase	608540	Severe psychomotor retardation, hypotonia, acquired microcephaly, intractable seizures, fever, coagulopathy, nephrotic syndrome, early death
CDG-Il	ALG9	Mannosyltransferase Dol-P-Man: Man_6- and Man_8-$GlcNAc_2$-P-P-Dol mannosyltransferase	608776	Severe microcephaly, hypotonia, seizures, hepatomegaly
CDG-IIa	MGAT2	GlcNAc transferase 2	212066	Mental retardation, dysmorphism, stereotypies, seizures
CDG-IIb	GLS1	Glucosidase I	606056	Dysmorphism, hypotonia, seizures, hepatomegaly, hepatic fibrosis; death at 2.5 months
CDG-IIc	SLC35C1/ FUCT1	GDP-fucose transporter	266265	Recurrent infections, persistent neutrophilia, mental retardation, microcephaly, hypotonia; normal transferrin
CDG-IId	B4GALT1	β1,4 galactosyltransferase	607091	Hypotonia (myopathy), spontaneous haemorrhage, Dandy–Walker malformation
CDG-IIe	COG7	Conserved oligomeric Golgi complex subunit 7	608779	Fatal in early infancy; dysmorphism, hypotonia, intractable seizures, hepatomegaly, progressive jaundice, recurrent infections, cardiac failure
CDG-IIf	SLC35A1	CMP-sialic acid transporter	605634	Thrombocytopaenia, no neurological symptoms; normal transferrin, abnormal platelet glycoproteins
CDG-II/COG1	COG1	Conserved oligomeric Golgi complex subunit 1	606973	Hypotonia, growth retardation, progressive microcephaly, hepatosplenomegaly, mild mental retardation
Mucolipidosis II and III	GNPTA	UDP-GlcNAc: lysosomal enzyme, GlcNAc-P transferase	252500	Coarsening features, organomegaly, joint stiffness, dysostosis, median neuropathy at the wrist; MLIII is less severe than MLII, which presents in infancy
Congenital dyserythropoietic anaemia (CDA II)	Unknown	Unknown	224100	Anaemia, jaundice, splenomegaly, gall bladder disease

CDG, congenital disorder of glycosylation; Dol, dolichol; Glc, glucose; GlcNAc, N-acetylglucosamine; Man, mannose.

O-Mannose-Linked Glycan Biosynthesis

1. Activitation of the POMT1–POMT2 mannosyltransferase complex, which uses dolichol-P-mannose to add α-mannose to serine (S) or threonine (T) residues.
2. POMGnT1 then uses UDP-GlcNAc (N-acetylglucosamine) to add a ß1,2GlcNAc residue to the mannose.
3. The O-mannose structure is completed by the addition of a ß1,4-galactose and 2,3-sialic acid residues. The specific transferases have not been identified for these steps. Glucuronic acid, sulphate and other branches can also be added.

Synthesis of a generic glycosaminoglycan chain

The linkage-region and chondroitin sulphate/dermatan sulphate and heparan sulphate/heparin structures are shown on Slide 24.

1. The common tetrasaccharide core linkage region is assembled
2. A specific N-acetylhexosamine transferase determines whether elongation proceeds towards chondroitin sulphate/dermatan sulphate (GalNAc (N-acetylgalactosamine) transferase) or heparan sulphate/heparin (GlcNAc transferase).
3. Separate co-polymerases then carry out stepwise additions of either glucuronic acid-GalNAc (chondroitin sulphate/dermatan sulphate) or glucuronic acid-GlcNAc (heparan sulphate/heparin).
4. These glycan backbones are modified in several ways that include epimerization of glucuronic acid to iduronic acid residues, N-de-acetylation/sulphation (heparan sulphate/heparin) and an array of 3′-phosphoadenosine-5′-phosphosulphate-dependent sulphation reactions generate enormous sequence diversity.

Human diseases caused by genetic defects in O-glycosylation and glycolipid synthesis pathways

Disorder	Gene	Enzyme	OMIM	Key Features
Defects in O-glycosylation pathways				
Walker–Warburg syndrome	POMT1/ POMT2	*O*-mannosyltransferase 1	236670	Type II lissencephaly, cerebellar malformations, ventriculomegaly, anterior chamber malformations, severe delay; death in infancy
Fukuyama muscular dystrophy	FCMD	Fukutin, a putative glycosyltransferase	253800	Cortical dysgenesis, myopia, weakness and hypotonia; 40% have seizures
Congenital muscular dystrophy type 1C (MDC1C)	FKRP	Fukutin-related protein, a putative glycosyltransferase	606612	Hypotonia, impaired motor development, respiratory muscle weakness
Congenital muscular dystrophy type 1D (MDC1D)	LARGE	Putative glycosyltransferase	608840	Muscular dystrophy with profound mental retardation
Hereditary inclusion body myopathy-II (IBM2)	GNE	UDP-GlcNAc epimerase/kinase	600737	Adult onset with progressive distal and proximal muscle weakness; spares quadriceps
Ehlers–Danlos syndrome	B4GALT7	β1,4-Galactosyltransferase 7	130070	Progeroid Ehlers-Danlos syndrome; macrocephaly, joint hyperextensibility
Hereditary multiple exostosis	EXT1/ EXT2	Glucuronyltransferase/GlcNAc transferase	133700	Multiple exostoses (diaphyseal, juxtaepiphyseal)
Chondrodysplasias	DTDST/ SLC26A2	Sulphate anion transporter	222600 600972 256050	Diastrophic dysplasia: airway collapse, early death in severe cases, adults reported. Achondrogenesis Ib: usually stillborn or early death of respiratory failure. Atelosteogenesis II: pulmonary hypoplasia, fatal in infants
Spondyloepimetaphyseal dysplasia	ATPSK2	3′-phosphoadenosine-5′-phosphosulphate synthase	603005	Abnormal skeletal development and linear growth
Macular corneal dystrophy types I and II	CHST6	Keratan sulphate 6-0-sulphotransferase	217800	Corneal clouding and erosions, painful photophobia
Familial tumoral calcinosis	GALNT3	GalNAc transferase	211900	Massive calcium deposits in skin and tissue
Tn syndrome	COSMC	Chaperone of β1,3GalT	230430	Anaemia, leukopaenia, thrombocytopaenia (somatic mutation)
Defects in glycolipid synthesis				
Paroxysmal nocturnal haemoglobinuria	PIGA	PI-GlcNAcT	311770	Complement-mediated haemolysis (somatic mutation)
Amish infantile epilepsy	SIAT9	Sia2,3Galβ1,4Glc-Cer synthase	609056	Tonic-clonic seizures, arrested development, neurological decline

CDG, congenital disorder of glycosylation; Cer, ceramide; Gal, galactose; GalNAc, N-acetylgalactosamine; Glc, glucose; GlcNAc, N-acetylglucosamine; Sia, sialic acid.

URLS

1. Glycosciences.de: http://www.dkfz-heidelberg.de/
2. KEGG GLYCAN: http://www.genome.jp/kegg/glycan/
3. Consortium for Functional Glycomics: http://www.functionalglycomics.org/
4. Bacterial Carbohydrate Structure DataBase: http://www.glyco.ac.ru/bcsdb3/
5. EuroCarbDB: http://www.eurocarbdb.org/
6. GlycomeDB: http://www.glycome-db.org/
7. GlycoCT: http://www.eurocarbdb.org/recommendations/encoding
8. GlycoMod: http://www.expasy.ch/tools/glycomod/
9. EuroCarbDB page of additional related links: http://www.eurocarbdb.org/links

References:

1. K.F. Aoki-Kinoshita, N. Ueda, H. Mamitsuka, and M. Kanehisa, ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains, *Bioinformatics*, 22, e25-e34, 2006.
2. A. Bohne-Lang, E. Lang, T. Forster, C.W. von der Lieth. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res*, 336: 1-11, 2001.
3. C.A. Cooper, E. Gasteiger, N.H. Packer, GlycoMod – A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1, 340-349, 2001.
4. H.H. Freeze, Genetic defects in the human glycome. *Nature Reviews Genetics*, 7, 537-551, 2006.
5. S. Hakomori. Glycosylation defining cancer malignancy: New wine in an old bottle. *PNAS*, 99(16), 10231-10233, 2002.
6. K. Hashimoto, K.F. Aoki-Kinoshita, N. Ueda, M. Kanehisa, and H. Mamitsuka, A New Efficient Probabilistic Model for Mining Labeled Ordered Trees, *Proc. KDD*, 177-186, 2006.
7. J. Hirabayashi, T. Hashidate, Y. Arata, N. Nishi, T. Nakamura, M. Hirashima, T. Urashima, T. Oka, M. Futai, W.E.G. Muller, F. Yagi and K. Kasai, Oligosaccharide specificity of galectins: a search by frontal affinity chromatography, *Biochimica et Biophysica Acta (BBA) - General Subjects,* 1572(2-3), 232-254, 2002.
8. Y. Hizukuri, Y. Yamanishi, O. Nakamura, F. Yagi, S. Goto, M. Kanehisa. Extraction of leukemia specific glycan motifs in humans by computational glycomics, *Carbohydrate Research*, 340(14), 2270-8, 2005
9. T. Kuboyama, K. Hirata, K.F. Aoki-Kinoshita, H. Kashima, and H. Yasuda. A Gram Distribution Kernel Applied to Glycan Classification and Motif Extraction, Genome Informatics, 17(2), 25-34, 2006.
10. A.H. Merry and C.L.R. Merry. Glycoscience finally comes of age. *EMBO Reports*, 6(10), 900-903, 2005.
11. K. Ohtsubo and J.D. Marth. Glycosylation in cellular mechanisms of health and disease. *Cell.* 126(5), 855-67, 2006.
12. H. Tang, Y. Mechref and M.V. Novotny, Automated interpretation of MS/MS spectra of oligosaccharides, *Bioinformatics*, 21, i431-i439, 2005.
13. N. Ueda, K.F. Aoki-Kinoshita, A. Yamaguchi, T. Akutsu, and H. Mamitsuka, A probabilistic model for mining labeled ordered trees: capturing patterns in carbohydrate sugar chains, *IEEE Transactions on Knowledge and Data Engineering,* 17(8), 1051-1064, Aug 2005.
14. A. Varki, Nothing in Glycobiology Makes Sensei, except in the Light of Evolution. *Cell*, 126, 841-845, 2006.
15. C.W. von der Lieth, T. Luetteke, F. Martin, The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochimica et Biophysica Acta, General Subjects*, 1760(4), 568-577, 2006.
16. D.B. Williams, Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum. *J Cell Sci.* 119(Pt 4):615-23, 2006.