An introduction to bioinformatics for glycomics research

Kiyoko F. Aoki-Kinoshita
Soka University, Japan
Introduction

Biological role of carbohydrates as information containing molecules
Glyco-science

Glycosylation by gene products

Glycome
Gene products, interactions with carbohydrate environment

Glycosidases
Glycosyltransferases
Carbohydrate-modifying enzymes
Lectins, CBPs, growth factors

Post-Genomics
Genomics

CDG=Congenital Disorder of Glycosylation
CBP=Carbohydrate Binding Protein
Introduction

Nomenclature of Carbohydrates
Glyceraldehyde, the simplest Aldose, contains one chiral carbon atom carrying four different substituents and has therefore two different enantiomers.

\[
\begin{align*}
\text{CHO} & \quad \text{CH}_2\text{OH} \\
\text{H} \quad \text{C} \quad \text{OH} & \\
\text{D-Glyceraldehyde} \\
\text{CHO} & \quad \text{CH}_2\text{OH} \\
\text{H} \quad \text{C} \quad \text{OH} & \\
\text{L-Glyceraldehyde}
\end{align*}
\]

Conventional representation of a carbon atom (e.g. C-2 of d-glucose) in the Fischer projection.
Some common and biologically important monosaccharides

- Glc (β-D-Glucose)
- Man (β-D-Mannose)
- Gal (β-D-Galactose)
- Fuc (β-L-Fucose)
- GlcNAc (β-N-Acetylglucosamine)
- GalNAc (β-N-Acetylgalactosamine)
- IdoA (β-D-Iduronic acid)
- Rib (β-D-Ribose)
Reference Database of Monosaccharides
http://www.monosaccharidedb.org

Exact query:

Enter a residue name following the rules for monosaccharide residue notation:
Submit Query  Reset

Or build residue name from pull down menus:

Anomeric: \( \alpha \) (alpha)  
Abs. Config: \( L \)  
Base Type: Fuc (Fucose)  
Ring Type: \( \beta \) (pyranose)  
Uronic Acid:  

Modifications:

Position: \( \beta \)  Type: fluoro (Fluori)  

Properties:

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>228</td>
</tr>
<tr>
<td>Name</td>
<td>( \alpha )-Fucp3fluoro</td>
</tr>
<tr>
<td>BaseType</td>
<td>Fuc (Fucose)</td>
</tr>
<tr>
<td>RootType</td>
<td>Gal</td>
</tr>
<tr>
<td>Size</td>
<td>6</td>
</tr>
<tr>
<td>Anomeric</td>
<td>( \alpha )</td>
</tr>
<tr>
<td>Abs. Config</td>
<td>( L )</td>
</tr>
<tr>
<td>Ring Type</td>
<td>( \beta )</td>
</tr>
<tr>
<td>Stereocode</td>
<td>22112</td>
</tr>
</tbody>
</table>

Modifications:

<table>
<thead>
<tr>
<th>Modification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>fluoro (Fluorine atom)</td>
</tr>
<tr>
<td>Position</td>
<td>( \beta )</td>
</tr>
</tbody>
</table>
Oligosaccharide description

- **Tree structures of monosaccharides and linkages**
- **Nodes** = sugars/monosaccharides
- **Edges** = bonds/linkages

Root node

\[
\begin{align*}
\text{a-D-Manp} & \quad (1-6) + \\
\text{b-D-Galp} & \quad (1-4) - \text{b-D-GlcpNAc} \quad (1-2) + \\
\text{b-D-Galp} & \quad (1-4) - \text{b-D-GlcpNAc} \quad (1-2) + \\
\text{a-D-Manp} & \quad (1-3) + \\
\text{b-D-Galp} & \quad (1-4) - \text{b-D-GlcpNAc} \quad (1-4) + \\
\end{align*}
\]
Introduction

Glyco-related pathways
Overview of glycan biosynthetic pathways.

Hudson H Freeze
2006 Jul;7(7):537-51
# Glycan classes: functions and biosynthesis

<table>
<thead>
<tr>
<th>Glycan Type</th>
<th>Linkage</th>
<th>Synthesis</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protein</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-Linked</td>
<td>GlcNAc-β-asparagine</td>
<td>Added in ER, modified in Golgi</td>
<td>Protein folding and stability; complex formation; signalling; cell–cell recognition</td>
</tr>
<tr>
<td>O-Linked</td>
<td>Mannose-α-serine/threonine</td>
<td>Begins in ER, completed in Golgi</td>
<td>Stability and function of dystrophin glycoprotein complex</td>
</tr>
<tr>
<td>Xylose-β-serine</td>
<td></td>
<td>Golgi</td>
<td>Mechanical cushioning; establishment of growth-factor and morphogen gradients</td>
</tr>
<tr>
<td>GalNAc-α-serine/threonine</td>
<td></td>
<td>Golgi</td>
<td>Lubrication; barrier against pathogens; leukocyte/lymphocyte trafficking</td>
</tr>
<tr>
<td><strong>Lipid</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycosphingolipid</td>
<td>Glucuronic acid-β-ceramide</td>
<td>Begins in ER, completed in Golgi</td>
<td>Lipid raft component; signalling; glycan-mediated cell–cell recognition</td>
</tr>
<tr>
<td>GPI anchor</td>
<td>N-glucuronic acid-inositol</td>
<td>Made in ER and transferred to proteins</td>
<td>Lipid raft component; haematopoeisis; protein trafficking</td>
</tr>
</tbody>
</table>

ER, endoplasmic reticulum; GalNAc, N-acetylgalactosamine; GlcNAc, N-acetylglucosamine

---

Hudson H Freeze  
Genetic defects in the human glycome.  
Nat Rev Genet.  
2006 Jul;7(7):537-51
N-linked glycan biosynthetic pathway

Steps in the pathway at which genetic disorders occur are indicated, with the associated genes underneath, as are steps at which an animal model is available. MPDU1 encodes a protein that enables the utilization of dolichol-P-mannose and dolichol-P glucose, but does not catalyse the reactions.

Hudson H Freeze
2006 Jul;7(7):537-51
Human diseases caused by genetic defects in N-glycosylation pathways

- **Congenital disorders of glycosylation** (19 distinct genes)
  - Mental retardation, seizures, epilepsy, ...

- **Mucolipidosis I & II**
  - Coarsening features, organomegaly, joint stiffness, ...

- **Congenital dyserythropoietic anaemia (CDA II)**
  - Anaemia, jaundice, splenomegaly, gall bladder disease
NS, 2S, 3S, 4S and 6S represent 2-N-, 2-O-, 3-O-, 4-O- and 6-O-sulphate, in that order.
Human diseases caused by genetic defects in O-glycosylation pathways

- Walker-Warburg syndrome
- Fukuyama muscular dystrophy
- Ehlers-Danlos syndrome
- Chondrodysplasias
- Macular corneal dystrophy
- Tn syndrome
- others

Human diseases caused by genetic defects in glycolipid synthesis

- Paroxysmal nocturnal haemoglobinuria
- Amish infantile epilepsy
Calnexin and calreticulin are related proteins that comprise an ER chaperone system that ensures the proper folding and quality control of newly synthesized glycoproteins.

Glyco-databases and data formats
Carbohydrate Structure Databases

- CarbBank
- SWEET-DB / glycosciences.de
- KEGG GLYCAN
- Consortium for Functional Glycomics
- BCSDB
- EuroCarbDB

Commercial databases:
- GlycoSuite (Proteome Systems, Ltd.)
- Glycomics DB (Glycominds, Ltd.)
CarbBank

- Developed by Complex Carbohydrate Research Center, University of Georgia
- Community database of carbohydrates
- Project ended due to lack of funding in 1996
GLYCOSCIENCES.de DB

- http://www.glycosciences.de
- Combines CarbBank and Sugabase using a common web-based interface
- Provides searching by bibliography, structure, NMR and MS, as well as by LINUCS ID
<table>
<thead>
<tr>
<th>a-D-Manp</th>
<th>1-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-D-Manp</td>
<td>1-4</td>
</tr>
<tr>
<td>a-D-GlcNac</td>
<td>1-3</td>
</tr>
</tbody>
</table>

- [ ] with 3D-Coo-ordinates [Sweet2]
- [ ] with residues
- [ ] min # residues
- [ ] with PDB entries
- min resolution
- [ ] chains
- [ ] all methods

**Structure Search**
You can enter from monosaccharide to pentasaccharides, please use the field in the center.

**Advanced mode**

Click [here](http://www.dkff.de/spec/dkff/databases/structure/structure_search) to reset input.
SWEETRDBb(3)

A-D-NEU5AC- (2-6) - B-D-GALP- (1-4) - B-D-GLCNAC- (1-2) - A-D-MANP- (1-6) +

A-D-NEU5AC- (2-6) - B-D-GALP- (1-4) - B-D-GLCNAC- (1-4) +

A-D-MANP- (1-3) +

A-D-NEU5AC- (2-6) - B-D-GALP- (1-4) - B-D-GLCNAC- (1-2) +

A-D-MANP- (1-6) +

A-D-MANP- (1-3) +

A-D-GLCNAC- (1-2) - A-D-MANP- (1-6) +

A-D-MANP- (1-4) - A-D-GLCNAC- (1-4) - B-D-GLCNAC- (1-4) - ASN

A-D-MANP- (1-3) +

B-D-GLCNAC- (1-2) - A-D-MANP- (1-6) +

B-D-MANP- (1-4) - A-D-GLCNAC- (1-4) - B-D-GLCNAC- (1-4) - ASN

A-D-MANP- (1-3) +
KEGG GLYCAN

- Based on CarbBank as well as input from scientists
- All data is linked with KEGG’s other resources: GENES, PATHWAY, KO and literary databases
- Several tools for analysis available
KEGG GLYCAN, a part of the KEGG LIGAND database, is a collection of experimentally determined glycan structures. It contains all unique structures taken from CarbBank, structures entered from recent publications, and structures present in KEGG pathways.

- DBGET search
- LIGAND relational database search
- Monosaccharide codes

Composite Structure Map (CSM) is a framework of all possible glycan structures generated from the KEGG GLYCAN database. CSM can be used to examine the structural repertoire inferred from genomic or transcriptomic repertoire of glycosyltransferase genes.

- KEGG GLYCAN composite structure map
- Glycosyltransferase reactions
- Glycosyltransferases
- KO groups for glycosyltransferases

http://www.genome.jp/kegg/glycan/
ORTHOLOGY: K03843

Entry | K03843  
---|---
Name | ALG2
Definition | alpha-1,3/alpha-1,6-mannosyltransferase
Class | Metabolism; Glycan Biosynthesis and Metabolism; W-Glycan biosynthesis [PATH:KO00510]
Protein Families; Metabolism; Glycosyltransferases [FR:KO01C03]

Other DBs
- FN: R05973 R06238
- EC: 2.4.1.132 2.4.1.-
- CAZY: GT4

Genes
- ESA: 85365 (Alg62)
- PFR: 472993 (LOC472993)
- MMU: 56737 (Alg2)
- FNO: 313231 (Alg2)
- CFA: 474780 (LOC474780)
- ETA: 538899 (MGC140299)
- XLA: 446622 (alg2-prov)
- XTR: 595052 (alg2)
- SPU: 589243 (LOC589243)
- DMX: Dmc1_cq1291
- CEL: F09B5.2
- ATH: AT1G7860C
- OSA: 4336813
- CMZ: CMT168C
- SCZ: YGL065C (Alg2)
- AGO: AG05_AFLC98W
- SPO: SPBC11B10.01
- ANI: AT0G374.2
- AFM: AFUA_5613210
- AOR: A0090120000461
- CMI: CH000660

Glyco: and Metabolism; N-Glycan
KO (KEGG Orthology) groups for glycosyltransferases

**N-glycan biosynthesis**

<table>
<thead>
<tr>
<th>KO</th>
<th>Human gene</th>
<th>Glycosidic linkage</th>
<th>Reaction</th>
<th>EC number</th>
<th>CAZy</th>
</tr>
</thead>
<tbody>
<tr>
<td>K01001 ALG7</td>
<td></td>
<td>GlcNAc - PP-Dol</td>
<td>R05969</td>
<td>2.7.8.15</td>
<td></td>
</tr>
<tr>
<td>K07432 ALG13 (sce)</td>
<td></td>
<td>GlcNAc b1-4 GlcNAc</td>
<td>R05970</td>
<td>2.4.1.141</td>
<td></td>
</tr>
<tr>
<td>K07441 ALG14 (sce)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K03842 ALG1</td>
<td></td>
<td>Man b1-4 GlcNAc</td>
<td>R05972</td>
<td>2.4.1.142</td>
<td>GT33</td>
</tr>
<tr>
<td>K03843 ALG2</td>
<td></td>
<td>Man a1-3 Man</td>
<td>R05973</td>
<td>2.4.1.132</td>
<td>GT4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man a1-6 Man</td>
<td>R06238</td>
<td>2.4.1.-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man a1-2 Man</td>
<td>R06127</td>
<td>2.4.1.-</td>
<td></td>
</tr>
<tr>
<td>K03844 ALG11 (sce)</td>
<td></td>
<td>Man a1-2 Man</td>
<td>R06128</td>
<td>2.4.1.-</td>
<td>GT4</td>
</tr>
<tr>
<td>K00721 DPM1</td>
<td></td>
<td>Man b1- P-Dol</td>
<td>R01009</td>
<td>2.4.1.83</td>
<td>GT2</td>
</tr>
<tr>
<td>K03845 ALG3</td>
<td></td>
<td>Man a1-3 Man</td>
<td>R06258</td>
<td>2.4.1.130</td>
<td>GT58</td>
</tr>
<tr>
<td>K03846 ALG9</td>
<td></td>
<td>Man a1-2 Man</td>
<td>R06259</td>
<td>2.4.1.130</td>
<td>GT22</td>
</tr>
<tr>
<td>K03847 ALG12</td>
<td></td>
<td>Man a1-6 Man</td>
<td>R06260</td>
<td>2.4.1.130</td>
<td>GT22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man a1-2 Man</td>
<td>R06261</td>
<td>2.4.1.130</td>
<td></td>
</tr>
<tr>
<td>K00729 ALG5</td>
<td></td>
<td>Glc b1- P-Dol</td>
<td>R01005</td>
<td>2.4.1.117</td>
<td>GT2</td>
</tr>
<tr>
<td>K03848 ALG6</td>
<td></td>
<td>Glc a1-3 Man</td>
<td>R06262</td>
<td>2.4.1.-</td>
<td>GT57</td>
</tr>
<tr>
<td>K03849 ALG8</td>
<td></td>
<td>Glc a1-3 Glc</td>
<td>R06263</td>
<td>2.4.1.-</td>
<td>GT57</td>
</tr>
<tr>
<td>K03850 ALG10</td>
<td></td>
<td>Glc a1-2 Glc</td>
<td>R06264</td>
<td>2.4.1.-</td>
<td>GT59</td>
</tr>
<tr>
<td>K00730 DDOST</td>
<td>RPN1</td>
<td>GlcNAc b1- Asn</td>
<td>R05976</td>
<td>2.4.1.119</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DDST2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KEGG GLYCAN Pathway Map

KEGG PATHWAY database is a collection of manually drawn KEGG pathway maps representing current knowledge on molecular interaction networks, including glycan biosynthesis and metabolism.

- Glycan biosynthesis and metabolism
- Overall relationship of pathway maps

KEGG GLYCAN Structure Map

KegDraw is a software tool for drawing chemical structures, glycan structures, and ISIS/Draw and ISIS/Draw Plus on Windows, Macintosh, and Linux, and is widely used in both academic and industrial applications.

- Download KegDraw

KCaM Search
**KEGG’s Glycan Biosynthesis and Metabolism Pathways**

<table>
<thead>
<tr>
<th>Glycan Pathways</th>
<th>Glycan Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Glycan biosynthesis</td>
<td>Glycosylphosphatidylinositol(GPI)</td>
</tr>
<tr>
<td>High-mannose type N-glycan biosynthesis</td>
<td>-anchor biosynthesis</td>
</tr>
<tr>
<td>N-Glycan degradation</td>
<td>Glycosphingolipid metabolism</td>
</tr>
<tr>
<td>O-Glycan biosynthesis</td>
<td>Blood group glycolipid</td>
</tr>
<tr>
<td>Chondroitin / heparan sulfate biosynthesis</td>
<td>biosynthesis - lactoseries</td>
</tr>
<tr>
<td>Keratan sulfate biosynthesis</td>
<td>Blood group glycolipid</td>
</tr>
<tr>
<td>Glycosaminoglycan degradation</td>
<td>biosynthesis - neo-lactoseries</td>
</tr>
<tr>
<td>Lipopolysaccharide biosynthesis</td>
<td>Globoside metabolism</td>
</tr>
<tr>
<td>Peptidoglycan biosynthesis</td>
<td>Ganglioside biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Glycan structures - biosynthesis 1</td>
</tr>
<tr>
<td></td>
<td>Glycan structures - biosynthesis 2</td>
</tr>
<tr>
<td></td>
<td>Glycan structures - degradation</td>
</tr>
</tbody>
</table>
KEGG GLYCAN Pathway Map

KEGG PATHWAY database is a collection of manually drawn KEGG pathway maps representing current knowledge on molecular interaction networks, including glycan biosynthesis and metabolism.

- Glycan biosynthesis and metabolism
- Overall relationship of pathway maps

KEgDraw

KEgDraw is a standalone Java application for drawing chemical compound structures and glycan structures in a way similar to ChemDraw and ISIS/Draw. KEgDraw runs on Mac, Windows, and Linux, and is made freely available to both academic and non-academic users.

- Download KEgDraw

KCaM Search

KEGG GLYCAN Structure Map
- DBGET search
- LIGAND relational database search
- Monosaccharide codes

**KEGG GLYCAN Pathway Map**

(Example) sce00510

**KEGG PATHWAY** database is a collection of manually drawn KEGG pathway maps representing current knowledge on molecular interaction networks, including glycan biosynthesis and metabolism.

- Glycan biosynthesis and metabolism
- Overall relationship of pathway maps

**KEGG GLYCAN Structure Map**

**KegDraw** is a standalone Java application for drawing chemical compound structures and glycan structures in a way similar to ChemDraw and ISIS/Draw. KegDraw runs on Mac, Windows, and Linux, and is made freely available to both academic and non-academic users.

- Download KegDraw

**KCaM Search**
**KEGG Glycan Search**

Enter query glycan: (in one of the three forms)

- **Glycan ID**: G00078 (Example) G00021

Select target database:
- KEGG GLYCAN
- CarbBank

Select program:
- Gapped (Approximate match)
- Ungapped (Exact match)

Select option:
- Global search
- Local search
<table>
<thead>
<tr>
<th>Entry</th>
<th>Structure</th>
<th>Name</th>
<th>Composition</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO0076</td>
<td>[Diag]</td>
<td>so-Neu5Ac-LacCer</td>
<td>([\text{Gal}]+[\text{GlcNAc}]_n) &amp; ([\text{GlcNAc}]_m) &amp; ([\text{Glc}]_l) &amp; ([\text{Cer}]_k)</td>
<td>Glycolipid; Sphingolipid</td>
</tr>
<tr>
<td>GO02772</td>
<td>[Diag]</td>
<td></td>
<td>([\text{Glc}]_4) &amp; ([\text{GlcNAc}]_3) &amp; ([\text{ Neu5Ac}]_1) &amp; ([\text{Cer}]_2)</td>
<td>Glycosylipid; Sphingolipid</td>
</tr>
<tr>
<td>GO4167</td>
<td>[Diag]</td>
<td></td>
<td>([\text{Glc}]_4) &amp; ([\text{GlcNAc}]_3) &amp; ([\text{ Neu5GaC}]_1) &amp; ([\text{Cer}]_1)</td>
<td>Glycosylipid; Sphingolipid</td>
</tr>
<tr>
<td>GO4168</td>
<td>[Diag]</td>
<td></td>
<td>([\text{Glc}]_4) &amp; ([\text{GlcNAc}]_3) &amp; ([\text{ Neu5Ac}]_1) &amp; ([\text{Cer}]_1)</td>
<td>Glycosylipid; Sphingolipid</td>
</tr>
<tr>
<td>Entry</td>
<td>Glycan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G04450</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Gal)4 (Glc)1 (GlcNAc)3 (LFuc)2 (Cer)1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1712.6 (Cer)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolipid; Sphingolipid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other DBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSD: 20753</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LinkDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>All DBs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KCF data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show</td>
</tr>
<tr>
<td>Entry</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>GO0078</td>
</tr>
<tr>
<td>GO00670</td>
</tr>
<tr>
<td>GO02772</td>
</tr>
<tr>
<td>GO4167</td>
</tr>
<tr>
<td>GO4168</td>
</tr>
</tbody>
</table>
Similarity-Score : 700

Query :

Entry : G04450
Consortium for Functional Glycomics (CFG)

- Consortium home page: http://www.functionalglycomics.org/
- Consortium of major universities and research institutes worldwide
- Aim: to provide a central resource for glycomics research
- Also provides requested resources to promote participating investigators’ research
  - Glycan arrays and data
  - Mass spectra analysis…
- CFG glycan database
Glycan Database

Updates
- First cut version of glycan structures database
- Contains nearly 7500 entries
- Each entry contains structural and chemical information as well as related references
- Different search interfaces are provided via the menu above
- The database will be regularly updated with newly synthesized or discovered glycans

Search for glycans
- **Sub-structure**
- Molecular weight
- Composition
- Linear nomenclature
- Use multiple search criteria

Source of glycan structures
- N- and O-linked glycans from CarbBank
- Glycominds Ltd., seed database
- N- and O-linked glycans identified in tissues and cells analyzed by the Analytical Glycotechnology Core (C)
- Glycans elaborated on the glycan array
- Glycans synthesized by the Carbohydrate Synthesis Core (D) and available as a resource

Glycan nomenclature
- Glycans are displayed in several formats for ease of use.
- The Consortium nomenclature for representing glycans can be found here.

© 2002-2005 Consortium for Functional Glycomics. All rights reserved.
Glycan Search

Sub-Structure Search

Create a new structure to search for

Create structure starting with the template

Create structure starting with the template

[To find glycan structures from the database containing specific sub-structures.]

© 2002-2005 Consortium for Functional Glycomics. All rights reserved.

This website is best viewed with Internet Explorer 5.5+ or Netscape 6.0+.
1. Click on monosaccharide above.
2. Select monosaccharide and linkage to add to the above selected.
3. Click submit to modify structure.
4. Click update query when done.

Add monosaccharide:
- Monosaccharide Linkage:
- GalNAc
- a
- Select Modifier:
- 1

Edit monosaccharide:
- Select monosaccharide to replace the selected: GalNAc
Glycan Search

Sub-Structure Search

Create a new structure to search for

Create structure starting with the template

Create structure starting with the template

Man<sub>a1</sub><sup>6</sup> Man<sub>a1</sub><sup>3</sup> Man<sub>b1</sub><sup>4</sup> GlcNAc<sub>b1</sub><sup>4</sup> GlcNAc

[To find glycan structures from the database containing specific sub-structures.]

Submit
<table>
<thead>
<tr>
<th>Oligosaccharide Molecular Wt.</th>
<th>IUPAC</th>
<th>Composition</th>
<th>Family</th>
<th>Sub Family</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2842.89</td>
<td>$\text{GalB1-4GlcNAcB1-2(\text{GalB1-4GlcNAcB1-4})}$</td>
<td>HEX$_2$DHEX$_2$HEXNAc$_6$</td>
<td>N-linked</td>
<td>Complex</td>
<td>Sus scrofa (Pig): Ovary: Follicle cells</td>
</tr>
<tr>
<td>2738.49</td>
<td>$\text{NeuAc2-3GalB1-4GlcNAcB1-2(\text{GalB1-4GlcNAcB1-4})}$</td>
<td>HEX$_2$DHEX$_1$HEXNAc$_5$NEUAc$_2$</td>
<td>N-linked</td>
<td>Complex</td>
<td>Cricetulus griseus (Chinese Hamster): CHO cells</td>
</tr>
<tr>
<td>5360.97</td>
<td>$\text{GlcNAcB1-3GalB1-4GlcNAcB1-3GlcNAcB1-4(\text{FucB1-6})}$</td>
<td>HEX$_1$DHEX$_2$HEXNAc$_6$</td>
<td>N-linked</td>
<td>Complex</td>
<td>Sus scrofa (Pig): Ovary: Follicle cells</td>
</tr>
<tr>
<td>1422.29</td>
<td>$\text{Mann1-3GalB1-4GlcNAcB1-2(\text{Mann1-6})}$</td>
<td>HEX$_3$DHEX$_2$HEXNAc$_4$</td>
<td>N-linked</td>
<td>Complex</td>
<td></td>
</tr>
</tbody>
</table>
Glycan: carbNlink_21394_P

Cartoon Representation

IUPAC 2D Representation

IUPAC Code

Gal b1-4 2% | Gal b1-4 1% | 2% | Gal b1-4 GlcNAc b1-2 (2%) | 1% | Gal b1-4 GlcNAc b1-4 | Man a1-3 (2%) | 1% | Gal b1-4 GlcNAc b1-2 (2%) | 1% | Gal b1-4 GlcNAc b1-6 | Man a1-5 | Man b1-4 GlcNAc b1-4 (Fuc a1-6) GlcNAc

Linear Code

Ab4=2% | Ab4=1% | 2% | Ab4=Nb2 | 2% | Ab4+Gb4 | Ma3 | 2% | Ab4+Nb2 | 2% | Ab4+Gb4+Ma6 |

General Information

Glycan Family: N-linked
Sub. Family: Complex
IUPAC Code

Gal b1-4 GlcNAc b1-4 Man a1-3 Man b1-4 GlcNAc b1-2(2% 1% Cal b1-4 GlcNAc b1-4) Man a1-3(2% 1% Cal b1-4 GlcNAc b1-2(2% 1% Cal b1-4 GlcNAc b1-6) Man a1-6) Man b1-4 GlcNAc b1-4(Fuc a1-6) GlcNAc

Linear Code

Ab4=2%|Ab4=1%|2%|1%Ab4GNa2(2%|1%Ab4GNa4)Ma3(2%|1%Ab4GNa2(2%|1%Ab4GNa6)Ma6) Mb4GNa4(Fa5)GNa

General Information

Glycan Family: N-linked
Sub. Family: Complex
Last Updated: 05/18/2004
Oligosaccharide Molecular Wt.: 2858.6091
Calculated Oligosaccharide Molecular Wt.: 2842.59
Per Methylated MW.: 3551
Composition: dHex₂ HexNAc₆ Hex₉
Status: Public

References


Biological Sources

<table>
<thead>
<tr>
<th>Taxonomy Name</th>
<th>Organ</th>
<th>Tissue Type</th>
<th>Cell Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sus scrofa(Pig)</td>
<td>Ovary</td>
<td></td>
<td>Follicle cells</td>
</tr>
</tbody>
</table>
BCSDB: Bacterial Carbohydrate Structure DataBase
http://www.glyco.ac.ru/bcsdb/start.shtml

- Provides structural, bibliographic, taxonomic and related information on bacterial carbohydrate structures.
- Data based on Carbbank and manual data posting (structures published after 1995, approx. 3000 records).
- >95% coverage of the scope of bacterial carbohydrates.
  - *Bacterial* = structure has been found in bacteria or obtained by modification of those found in bacteria.
  - *Carbohydrate* = structure composed of any residues linked by glycosidic, ester, amidic, ketal, phospho- or sulpho-diester bonds, in which at least one residue is a sugar or its derivative.
- Each record includes structure, bibliography, abstract, keywords, biological source, methods used to elucidate the structure, bioactivity, NMR assignment tables, etc.
- Search by IDs, bibliographic data and keywords, biological source, the fragment of structure and NMR data.
- Data cross-linked with GlycoSCIENCES.DB
BCSDB

Bacterial Carbohydrate Structure DataBase

currently 8341 structures.
last update: 2007 Feb 23; latest publication: 2005

Search using:  
- BCSDB ID
- Bibliography
- (Sub)structure
- Microorganism
- NMR signals

Submit data:  
- In a form
- In a file

Help:  
- About
- Usage
- NMR prediction
- Structure encoding
- Monomer namespace
- For programmers
- Credits
- Feedback

Admin:  
- Maintenance
- Data export
1. (BCSDB ID: 114464)

Akiba S, Yamamoto K, Kumagai H
Effects of size of carbohydrate chain on protease digestion of Aspergillus niger endo-α-1,4-glucanase
Bioscience, Biotechnology, and Biochemistry 59 (1995) 1048-1051

```
\text{aDMan}_p \text{(1-6)} + \\
\text{aDMan}_p \text{(1-3)} \text{aDMan}_p \text{(1-6)} + \\
\text{aDMan}_p \text{(1-2)} \text{aDMan}_p \text{(1-3)} \text{bDMan}_p \text{(1-4)} \text{bDGlcpNAc} \text{(1-4)} \text{bDGlcpNAc} \text{(1-4)} \text{Asn}
```

*Aspergillus niger* IFO31125 (NCBI Taxonomy)

2. (BCSDB ID: 129096)

Altmann F, Schweiszer S, Weber C
Kinetic comparison of peptide N-glycosidases F and A reveals several differences in substrate specificity
Glycoconjugate Journal 12 (1995) 84-93

```
\text{aDMan}_p \text{(1-6)} + \\
\text{aDMan}_p \text{(1-3)} \text{aDMan}_p \text{(1-6)} + \\
\text{aDMan}_p \text{(1-3)} \text{bDMan}_p \text{(1-4)} \text{bDGlcpNAc} \text{(1-4)} \text{bDGlcpNAc} \text{(1-4)} + \\
\text{Val} \text{(1-2)} \text{Ser} \text{(1-2)} \text{Asn} \text{(1-2)} \text{Tyr} \text{(1-2)} \text{Ser} \text{(1-2)} \text{Ile} \text{(1-2)} \text{Asp} \text{(1-2)} \text{Gly}
```

*Aspergillus oryzae* (NCBI Taxonomy)
Aspergillus niger IPO31125 (NCBI Taxonomy)

Taxonomic group: fungi (Peyling: Ascomycota)

Structure type: oligomer
Compound class: N-linked glycoprotein

Comments, role: Parent molecule: endo-α-1,4-glucanase

NCBI Taxonomy ref: TaxIDs: 5061

Make GLYDE 1.2 description

Predict $^{13}$C NMR assignment table

Find this structure in GlycoSClENCES DB

Collapse this record
**GlycoSCIENCES data retrieval**

This page displays GlycoSCIENCES DB data for the structure you specified.

**Structure** (LinusID=1232). To go to the corresponding Glycosciences DB page click here.

![Structure diagram]

Molecular weight: 1495  
Chemical formula: C_{66}H_{94}N_{4}O_{42}

**^1H NMR data:**  
Recorded in 300 at 0.0 K  
Spectrometer: 400.0 MHz

<table>
<thead>
<tr>
<th>Residue</th>
<th>Linkage</th>
<th>Atom</th>
<th>δ, ppm</th>
<th>J, from</th>
<th>J, to</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-D-GlcNAc</td>
<td>4</td>
<td>H1</td>
<td>5.04</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>b-D-GlcNAc</td>
<td>4</td>
<td>NAC</td>
<td>2.01</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>b-D-GlcNAc</td>
<td>4</td>
<td>NAC</td>
<td>2.01</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>b-D-GlcNAc</td>
<td>4,4</td>
<td>H1</td>
<td>4.60</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>b-D-GlcNAc</td>
<td>4,4</td>
<td>NAC</td>
<td>2.06</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>b-D-Manp</td>
<td>4,4,4</td>
<td>H1</td>
<td>4.77</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>b-D-Manp</td>
<td>4,4,4</td>
<td>H2</td>
<td>4.23</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>6,4,4,4</td>
<td>H1</td>
<td>4.87</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>6,4,4,4</td>
<td>H2</td>
<td>4.14</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>6,6,4,4</td>
<td>H1</td>
<td>4.91</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>6,6,4,4</td>
<td>H2</td>
<td>3.98</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>3,6,4,4</td>
<td>H1</td>
<td>5.10</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>a-D-Manp</td>
<td>3,6,4,4</td>
<td>H2</td>
<td>4.07</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>
EuroCarbDB – Design Study

- http://www.eurocarbdb.org/
- Based in Europe, but participants from universities and research groups worldwide
- Distributed infrastructure to integrate multiple resources with a single interface
Data Modeling

◆ Foremost issue in handling glycan structures for comparison and analysis

◆ A few models/formats currently available:
  • LINUCS
  • KCF
  • Linear Code©
  • GLYDE (XML)
  • GlycoCT
Glycome informatics

- Glycome: the repertoire of glycans in a cell, tissue, or organism
- Glycome informatics: Algorithms, methods and computational models for the study of the glycome
Current glycome informatics

- Glycomics:
  - Automated mass spectrometry annotation
- Computer-theoretic algorithms for tree alignments
- Probabilistic models (mining) for patterns in glycans
- Kernel methods for glycan classification
Glycomics Techniques

- Mass spectrometry of glycoproteins: prediction/annotation
  - Mizuno et al., Anal. Chem, 1999
  - GlycoMod (Cooper et al, Proteomics, 2001)
  - StrOligo (M. Ethier et al, Methods Mol Biol., 2006)
  - Cartoonist (D. Goldberg et al, Proteomics, 2005)
  - Glyco-Peakfinder (K. Maas, R. Ranzinger et al, Proteomics, 2007)
  - GlycoWorkbench (A. Ceorni et al., 2007)
  - GLYCH (H. Tang et al, Bioinformatics, 2005)
Automated Annotation of Mass Spectrometry Data
GlycoMod

- Predicts the possible oligosaccharide structures that occur on proteins from their experimentally determined masses.
- Can be used for free or derivatized oligosaccharides and for glycopeptides
Experimental workflow for (semi-)automatic determination of glycan structures from raw data to fully assigned spectrum via composition analysis (GlycoPeakFinder) and fragment matching (GlycoWorkbench).
Nomenclature of MS fragments of carbohydrates as defined by Domon and Costello
GlycoWorkbench
MS: Annotation of fragments

http://www.eurocarbdb.org/applications
GlycoWorkbench
MS: Annotation of fragments

http://www.eurocarbdb.org/applications
Current glycome informatics

- Automated mass spectrometry annotation
- Computer-theoretic algorithms for tree alignments
- Probabilistic models (mining) for patterns in glycans
- Kernel methods for glycan classification
Computer Theoretic Techniques

- KCaM: K.F. Aoki et al., NAR, 2004
- Score matrix for glycan linkages, K.F. Aoki et al., Bioinformatics, 2005
- Least common super tree approximation algorithm for reconstructing glycans from spectral data, K.F. Aoki-Kinoshita et al., ISAAC 2006
Glycan structure comparison

- Calculating glycan “similarity”
  - Efficiency
  - Biologically meaningful
- Data mining techniques
- Prediction:
  - In layman’s terms: determining whether or not a given glycan belongs to a particular class
Glycan structure comparison: KCaM

- KEGG Carbohydrate Matcher
- Glycan alignment tool for KEGG GLYCAN
- Maximum Common Subtree algorithm
- Dynamic programming approach
  - Smith-Waterman
  - Needleman-Wunsch
KCaM: KEGG Carbohydrate Matcher

- Smith-Waterman sequence alignment algorithm (global and local)

\[
\begin{align*}
S[i, 0] &= d \cdot i, \\
S[0, j] &= d \cdot j, \\
S[i, j] &= \max \left\{ S[i, j - 1] + d, S[i - 1, j] + d, S[i - 1, j - 1] + w(x_i, y_j) \right\} \\
S[i, 0] &= 0, \\
S[0, j] &= 0,
\end{align*}
\]
KCaM: KEGG Carbohydrate Matcher

- Maximum Common Subtree Algorithm

\[ R[u, 0] = 0, \]
\[ R[0, v] = 0, \]
\[ R[u, v] = 1 + \max_{\psi \in M(u, v)} \left\{ \sum_{u_\bar{v} \in \text{sons}(u)} R[u_\bar{v}, \psi(u_\bar{v})] \right\} \]
Glycan Score Matrix

- Like PAM or BLOSUM for proteins
- Improved KCaM using score matrix
- Similarity measures of matrix components (glycan components)
- Statistical insight into glycan composition
Method

- Matrix entries: “link” = monosaccharides + bond type
- “Families” determined by hierarchically clustering KEGG GLYCAN based on KCaM similarity scores
- Calculations performed similar to BLOSUM matrix for protein sequences
## Improved alignments

<table>
<thead>
<tr>
<th>Without Matrix Glycan</th>
<th>Score</th>
<th>With Matrix Glycan</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>G00086</td>
<td>8.0</td>
<td>G04134 *</td>
<td>5.90797</td>
</tr>
<tr>
<td>G00192</td>
<td>8.0</td>
<td>G04072 *</td>
<td>5.54569</td>
</tr>
<tr>
<td>G04134 *</td>
<td>7.0</td>
<td>G05073 *</td>
<td>5.20216</td>
</tr>
<tr>
<td>G04906 *</td>
<td>7.0</td>
<td>G04906 *</td>
<td>5.09453</td>
</tr>
<tr>
<td>G00407 *</td>
<td>6.0</td>
<td>G05305 *</td>
<td>4.99696</td>
</tr>
<tr>
<td>G00975</td>
<td>6.0</td>
<td>G04140 *</td>
<td>4.9072</td>
</tr>
</tbody>
</table>

**Query**

![Query Diagram](image)

**G00086**

![G00086 Diagram](image)

**G04134**

![G04134 Diagram](image)
### Individual Matrix Entries

<table>
<thead>
<tr>
<th>Aligned Linkage Child</th>
<th>Aligned Linkage Parent</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuc1, a6GlcNAc</td>
<td>Fuc1, a6GlcNAc</td>
<td>2.45254</td>
</tr>
<tr>
<td>GlcNAc1, b4GlcNAc</td>
<td>GlcNAc1, b4GlcNAc</td>
<td>2.37549</td>
</tr>
<tr>
<td>Man1, b4GlcNAc</td>
<td>Man1, b4GlcNAc</td>
<td>2.32516</td>
</tr>
<tr>
<td>Glc1, b4GlcNAc</td>
<td>Glc1, b4GlcNAc</td>
<td>2.08472</td>
</tr>
<tr>
<td>Man1, a4Glc</td>
<td>Man1, a6Glc</td>
<td>2.03222</td>
</tr>
<tr>
<td>Man1, a3Glc</td>
<td>Man1, b3Glc</td>
<td>1.9977</td>
</tr>
<tr>
<td>Glc1, a2Glc</td>
<td>Glc1, a2Glc</td>
<td>1.99001</td>
</tr>
<tr>
<td>Glc1, a3Glc</td>
<td>Glc1, a3Glc</td>
<td>1.98493</td>
</tr>
<tr>
<td>GlcNAc1, b6GalNAc</td>
<td>GlcNAc1, b6GalNAc</td>
<td>1.96005</td>
</tr>
</tbody>
</table>
Current glycome informatics

- Automated mass spectrometry annotation
- Computer-theoretic algorithms for tree alignments
- Probabilistic models (mining) for patterns in glycans
- Kernel methods for glycan classification
Mining in Glycome Informatics

◆ Probabilistic Models
  • PSTMM, N. Ueda et al, TKDE, 2005
  • Profile PSTMM, K.F. Aoki-Kinoshita et al, ISMB 2006
  • OTMM, Hashimoto et al, KDD 2006

◆ Previous work on probabilistic trees
  • Hidden Tree Markov Model, HTMM (Diligenti et al., 2003) for image classification
HTMM Cannot Capture Sibling Dependencies!
Probabilistic Sibling Tree Markov Model (PSTMM)
Inference and learning

- Estimating the parameters:
  - To “learn” patterns found in given data

- Calculating the likelihood of a set of trees:
  - To determine which data are considered to belong to the same class as learned data

- Finding the most likely state transition:
  - To retrieve the learned patterns
  - To apply to multiple tree alignments
Learned Classification

High Mannose

Hybrid

Complex
Summary of PSTMM Results

- There indeed seem to exist sibling-dependent relationships in glycans!
- Statistical analysis of glycans seem appropriate considering the noisiness of the data
  - Prediction of missing information
  - Further classification groups based on patterns found within a class of glycans
Provided binding affinity data for a specific lectin, compute the most likely structure being recognized.

Statistically compute the key patterns of sulfation in GAGs based on various biological measurements (i.e. inhibition).
Glycan recognition

- Glycans are modified, degraded, recognized by various types of proteins
  - Much research focuses on understanding the structure of the lectins that bind to glycans
  - Recognition of the substructures at the leaves
Lectin-glycan experiment

- Many classes of lectins (glycan-binding proteins)
  - Recognize specific monosaccharides at the leaves
- Galectins recognize Galactose residues
- FAC analysis has enabled high-throughput binding affinity analysis of galectins and glycans (J. Hirabayashi et al, 2002)


Lectin-glycan experiment

<table>
<thead>
<tr>
<th></th>
<th>Gal-3 affinity (weight)</th>
<th>Gal-9N affinity (weight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA3</td>
<td>1.28205 (1)</td>
<td>2.6316 (2)</td>
</tr>
<tr>
<td>fuc. NA3</td>
<td>1.21951 (1)</td>
<td>2.2222 (2)</td>
</tr>
<tr>
<td>NA3 type1</td>
<td>1.08696 (1)</td>
<td>1.6949 (0)</td>
</tr>
<tr>
<td>NA4</td>
<td>1.44928 (1)</td>
<td>5.5556 (5)</td>
</tr>
<tr>
<td>fuc. NA4</td>
<td>1.40845 (1)</td>
<td>4.3478 (4)</td>
</tr>
<tr>
<td>Galili penta.</td>
<td>1.47059 (1)</td>
<td>0.2273 (0)</td>
</tr>
<tr>
<td>Forssman penta.</td>
<td>0.16129 (0)</td>
<td>11.111 (11)</td>
</tr>
<tr>
<td>A-hexa</td>
<td>1.5873 (1)</td>
<td>3.8462 (3)</td>
</tr>
<tr>
<td>LN3</td>
<td>2.85714 (2)</td>
<td>1.2346 (0)</td>
</tr>
<tr>
<td>LN5</td>
<td>5.26316 (5)</td>
<td>8.3333 (8)</td>
</tr>
</tbody>
</table>
Lectin-binding glycan profiles

Gal-3

Accuracy 0.847
Precision 1.0
AUC 0.930

Gal-9N

Accuracy 0.910
Precision 0.918
AUC 0.931
Current glycome informatics

- Automated mass spectrometry annotation
- Computer-theoretic algorithms for tree alignments
- Probabilistic models (mining) for patterns in glycans
- Kernel methods for glycan classification
Kernel Methods

- Machine learning method
  - e.g. Support Vector Machines (SVM)
- Can handle features in high-dimensions
  - e.g. Expression data, pathway information, localization information, etc.
- Statistically computes commonalities by reducing the dimensions of the data
  - Data classification
  - Feature extraction

http://www-kairo.csce.kyushu-u.ac.jp/~norikazu/research.en.html
Leukemia-specific features


- Used KEGG GLYCAN data:
  - Entries whose CarbBank annotations were related to leukemic cells, erythrocytes, plasma and serum
  - Predicted possible glycan markers
  - Correlated well with experimental data

- Assessed CarbBank data and retrieved leukemia-specific glycans via annotations

- Found that glycan substructures of three residues (trimers) produced best accuracy

- Also used the fact that structures at the leaves should be distinguished from those at the root
Leukemia Kernel

Layer-specific trimers for each glycan

Hizukuri et al., Carbohydrate Research, 2005.
Leukemia Kernel

- A vector of all possible trimers $n$ where $x_n$ is the number of times trimer $x$ appears in a particular glycan $G = G(x_1, x_2, \ldots, x_n)$
- Glycans $X$ and $Y$ are compared by the following function:

$$
sim(X, Y) = \sum_{k=1}^{260} w_k x_k y_k,
$$

where $w_k$ is defined as

$$
w_k = 1 - \exp(-\alpha h) \quad \text{if } h > 1,
$$
$$
w_k = 1 \quad \text{if } h = 1,
$$

where $h$ is the layer of the matching substructures and $\alpha$ is a positive constant (in this work, the parameter is set to 0.5). When the matching substructure is found at the root, the weight is set to 1.
Leukemia Markers

- Supported experimental results

<table>
<thead>
<tr>
<th>Substructures</th>
<th>Layer</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha\text{-d-Neup5Ac-(2→3)}\beta\text{-d-Galp-(1→4)}\text{-GlcNAc}$</td>
<td>5</td>
<td>161.2</td>
</tr>
<tr>
<td>$\beta\text{-d-Galp-(1→4)}\beta\text{-d-GlcNAc-(1→2)}\text{-Manp}$</td>
<td>4</td>
<td>159.6</td>
</tr>
<tr>
<td>$\alpha\text{-d-Neup5Ac-(2→6)}\beta\text{-d-Galp-(1→4)}\text{-GlcNAc}$</td>
<td>5</td>
<td>148.8</td>
</tr>
<tr>
<td>$\beta\text{-GlcNAc-(1→2)}\alpha\text{-d-Manp-(1→3)}\text{-Manp}$</td>
<td>3</td>
<td>78.7</td>
</tr>
<tr>
<td>$\beta\text{-GlcNAc-(1→2)}\alpha\text{-d-Manp-(1→6)}\text{-Manp}$</td>
<td>3</td>
<td>77.6</td>
</tr>
</tbody>
</table>
Gram distribution kernel

- Kuboyama et al., Genome Informatics, 2006.
- Took the distribution of dimers, trimers, quatrimers, etc. to represent a glycan
- Able to extract features of any size
- Used the concept of q-grams
Q-gram
Gram distribution kernel

- Possible to count all q-grams for rooted ordered trees in linear time (Kuboyama et al., LLLL 2006)
- By calculating the distribution of q-grams in a tree, this kernel is able to capture more information, including a variety of q for various path lengths
- To verify the performance of the gram distribution kernel, used the same data set as used for testing the Layered-Trimer Kernel
- Also tested a data set of glycans related to the keywords “cystic fibrosis,” “bronchial mucin,” and “respiratory mucin”
Results: Features extracted

<table>
<thead>
<tr>
<th>Score</th>
<th>Substructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>226</td>
<td>(leaf)NeuAc2–α3Gal–β4</td>
</tr>
<tr>
<td>201</td>
<td>(leaf)NeuAc2–α6Gal–β4</td>
</tr>
<tr>
<td>201</td>
<td>(leaf)NeuAc2–α3Gal–β4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Substructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>–Gal–β4GlcNAc–β3Gal–β4</td>
</tr>
<tr>
<td>86</td>
<td>(leaf)Fuc–α2Gal–β4GlcNAc–β3</td>
</tr>
<tr>
<td>86</td>
<td>–GlcNAc–β3Gal–β4Glc–β1(root)</td>
</tr>
<tr>
<td>82</td>
<td>–Gal–β4GlcNAc–β3Gal–β4GlcNAc–β3</td>
</tr>
<tr>
<td>81</td>
<td>(leaf)Fuc–α2Gal–β4GlcNAc–β3Gal–β4</td>
</tr>
</tbody>
</table>
Results: performance

- Gram distribution vs. Leukemia kernel (layered trimer kernel)
Results: marker size
Results: marker size
Systems approach to unveiling structure-function relationship
Glycan synthesis is a non-template driven process. We can never be sure that the complete structural space of glycans is represented in the databases.

Theoretical Number of Isomers = \( E^n \times 2^{n_{\text{anomer}}} \times 2^{n_{\text{conf}}} \times (4^{n-1}) \)

<table>
<thead>
<tr>
<th>Monosaccharide</th>
<th>1</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disaccharide</td>
<td>2</td>
<td>256</td>
</tr>
<tr>
<td>Trisaccharide</td>
<td>3</td>
<td>27,648</td>
</tr>
<tr>
<td>Tetrasaccharide</td>
<td>4</td>
<td>4,194,304</td>
</tr>
<tr>
<td>Pentasaccharide</td>
<td>5</td>
<td>819,200,00</td>
</tr>
<tr>
<td>Hexasaccharide</td>
<td>6</td>
<td>195,689,447,42</td>
</tr>
</tbody>
</table>

Which glycan structures really exist in certain species? What do the databases say?
## Occurrence of monosaccharide residues (CarbBank nomenclature)

<table>
<thead>
<tr>
<th>Monosaccharide name</th>
<th>mammalian #</th>
<th>mammalian [%]</th>
<th>Mammalian human #</th>
<th>human [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-D-GLCPNAC</td>
<td>7319</td>
<td>26,1%</td>
<td>4705</td>
<td>26,69%</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>6389</td>
<td>22,8%</td>
<td>4178</td>
<td>23,70%</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>3659</td>
<td>13,1%</td>
<td>2073</td>
<td>11,76%</td>
</tr>
<tr>
<td>A-D-NEUP5AC</td>
<td>2101</td>
<td>7,5%</td>
<td>1465</td>
<td>8,31%</td>
</tr>
<tr>
<td>A-L-FUCP</td>
<td>1971</td>
<td>7,0%</td>
<td>1461</td>
<td>8,29%</td>
</tr>
<tr>
<td>B-D-MANP</td>
<td>1486</td>
<td>5,3%</td>
<td>900</td>
<td>5,10%</td>
</tr>
<tr>
<td>D-GLCNAC</td>
<td>675</td>
<td>2,4%</td>
<td>403</td>
<td>2,29%</td>
</tr>
<tr>
<td>D-GLCNAC-OL</td>
<td>598</td>
<td>2,1%</td>
<td>399</td>
<td>2,26%</td>
</tr>
<tr>
<td>D-GALNAC-OL</td>
<td>511</td>
<td>1,8%</td>
<td>355</td>
<td>2,01%</td>
</tr>
<tr>
<td>B-D-GLCP</td>
<td>423</td>
<td>1,5%</td>
<td>244</td>
<td>1,38%</td>
</tr>
<tr>
<td>B-D-GALPNAC</td>
<td>431</td>
<td>1,5%</td>
<td>230</td>
<td>1,30%</td>
</tr>
<tr>
<td>SULFATE</td>
<td>450</td>
<td>1,6%</td>
<td>198</td>
<td>1,12%</td>
</tr>
<tr>
<td>A-D-GALPNAC</td>
<td>248</td>
<td>0,9%</td>
<td>171</td>
<td>0,97%</td>
</tr>
<tr>
<td>D-GLC</td>
<td>197</td>
<td>0,7%</td>
<td>151</td>
<td>0,86%</td>
</tr>
<tr>
<td>5A-D-GALP</td>
<td>287</td>
<td>1,0%</td>
<td>103</td>
<td>0,58%</td>
</tr>
<tr>
<td>D-GALNAC</td>
<td>116</td>
<td>0,4%</td>
<td>91</td>
<td>0,52%</td>
</tr>
<tr>
<td>A-D-GLCP</td>
<td>161</td>
<td>0,6%</td>
<td>68</td>
<td>0,39%</td>
</tr>
<tr>
<td>B-D-GLCPA</td>
<td>94</td>
<td>0,3%</td>
<td>54</td>
<td>0,31%</td>
</tr>
<tr>
<td>D-GAL</td>
<td>56</td>
<td>0,2%</td>
<td>37</td>
<td>0,21%</td>
</tr>
<tr>
<td>D-GLC-OL</td>
<td>37</td>
<td>0,1%</td>
<td>34</td>
<td>0,19%</td>
</tr>
<tr>
<td>A-D-GLCPNAC</td>
<td>89</td>
<td>0,3%</td>
<td>31</td>
<td>0,18%</td>
</tr>
<tr>
<td>D-GAL-OL</td>
<td>38</td>
<td>0,1%</td>
<td>27</td>
<td>0,15%</td>
</tr>
<tr>
<td>D-GLCPNAC</td>
<td>37</td>
<td>0,1%</td>
<td>17</td>
<td>0,10%</td>
</tr>
<tr>
<td>A-D-NEUP5GC</td>
<td>132</td>
<td>0,5%</td>
<td>16</td>
<td>0,09%</td>
</tr>
<tr>
<td>B-D-XYL P</td>
<td>23</td>
<td>0,1%</td>
<td>13</td>
<td>0,07%</td>
</tr>
<tr>
<td>D-GALP</td>
<td>22</td>
<td>0,1%</td>
<td>12</td>
<td>0,07%</td>
</tr>
<tr>
<td>A-L-4-EN-THRHEXPA</td>
<td>40</td>
<td>0,1%</td>
<td>12</td>
<td>0,07%</td>
</tr>
<tr>
<td>?-D-GALPNAC</td>
<td>15</td>
<td>0,1%</td>
<td>11</td>
<td>0,06%</td>
</tr>
<tr>
<td>P</td>
<td>20</td>
<td>0,1%</td>
<td>11</td>
<td>0,06%</td>
</tr>
<tr>
<td>D-2,5-ANHYDRO-MAN-OL</td>
<td>13</td>
<td>0,0%</td>
<td>9</td>
<td>0,05%</td>
</tr>
</tbody>
</table>

**Mammalian: 5339  Human: 2128**

**Total number of different residues**

**Mammalian: 86  Human: 83**

**Stephan Herget / Rene Ranzinger**
### Occurrence of disaccharide residues (CarbBank nomenclature)

<table>
<thead>
<tr>
<th>Parent</th>
<th>from</th>
<th>to</th>
<th>Child</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-GALP</td>
<td>2837</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>2</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>1382</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>3</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>860</td>
</tr>
<tr>
<td>B-D-MANP</td>
<td>6</td>
<td>1</td>
<td>A-D-MANP</td>
<td>776</td>
</tr>
<tr>
<td>B-D-MANP</td>
<td>3</td>
<td>1</td>
<td>A-D-MANP</td>
<td>771</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>3</td>
<td>2</td>
<td>A-D-NEUP-5AC</td>
<td>742</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-MANP</td>
<td>732</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>6</td>
<td>2</td>
<td>A-D-NEUP-5AC</td>
<td>467</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>2</td>
<td>1</td>
<td>A-L-FUCP</td>
<td>436</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>3</td>
<td>1</td>
<td>A-L-FUCP</td>
<td>418</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>4</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>340</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>3</td>
<td>1</td>
<td>B-D-GALP</td>
<td>300</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>6</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>255</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-GALP</td>
<td>219</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>6</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>186</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>175</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>2</td>
<td>1</td>
<td>A-D-MANP</td>
<td>156</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>3</td>
<td>1</td>
<td>A-D-GLCP-2NAC</td>
<td>119</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>4</td>
<td>1</td>
<td>A-L-FUCP</td>
<td>117</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>110</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>3</td>
<td>1</td>
<td>A-D-MANP</td>
<td>92</td>
</tr>
<tr>
<td>A-D-MANP</td>
<td>6</td>
<td>1</td>
<td>A-D-MANP</td>
<td>88</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>6</td>
<td>1</td>
<td>A-L-FUCP</td>
<td>86</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>4</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>78</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>3</td>
<td>1</td>
<td>A-D-GALP</td>
<td>68</td>
</tr>
<tr>
<td>B-D-GALP</td>
<td>4</td>
<td>1</td>
<td>B-D-GALP-2NAC</td>
<td>62</td>
</tr>
<tr>
<td>B-D-GLCP-2NAC</td>
<td>3</td>
<td>1</td>
<td>B-D-GALP</td>
<td>45</td>
</tr>
<tr>
<td>A-D-GLCP-2NAC</td>
<td>3</td>
<td>1</td>
<td>B-D-GLCP-2NAC</td>
<td>39</td>
</tr>
<tr>
<td>A-D-NEUP-5AC</td>
<td>8</td>
<td>2</td>
<td>A-D-NEUP-5AC</td>
<td>31</td>
</tr>
</tbody>
</table>

**Human:** 2128

**Total number of different Disaccharide**

- **Human:** 171
  - once: 65
  - twice: 20
  - Three Times: 10

**Occurrence:**

- **71.7%**
- **89.9%**
- **95.5%**

Stephan Herget / Rene Ranzinger
Topologies of Glycans

Size of Glycan (Residues)

Number of Branching points

Stephan Herget / Rene Ranzinger
Mathematical Modelling to explore the structural space of glycan using Information from carbohydrate active enzymes

A Mathematical Model of N-Linked Glycosylation

Enzyme reaction rule tables to model reaction networks:
Parameters: spatial distribution of enzymes, transport, reaction kinetics, donor concentrations.

<table>
<thead>
<tr>
<th>Enzymes included</th>
<th>EC No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviations</td>
<td></td>
</tr>
<tr>
<td>ManI</td>
<td>3.2.1.113</td>
</tr>
<tr>
<td>ManII</td>
<td>3.2.1.114</td>
</tr>
<tr>
<td>FucT</td>
<td>2.4.1.68</td>
</tr>
<tr>
<td>GnTI</td>
<td>2.4.1.101</td>
</tr>
<tr>
<td>GnTII</td>
<td>2.4.1.143</td>
</tr>
<tr>
<td>GnTIII</td>
<td>2.4.1.144</td>
</tr>
<tr>
<td>GnTIV</td>
<td>2.4.1.145</td>
</tr>
<tr>
<td>GnTV</td>
<td>2.4.1.155</td>
</tr>
<tr>
<td>GnTE</td>
<td>2.4.1.149</td>
</tr>
<tr>
<td>GalT</td>
<td>2.4.1.38</td>
</tr>
<tr>
<td>SiaT</td>
<td>2.4.99.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Glycoform description scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man Number of mannose residues</td>
</tr>
<tr>
<td>Fuc Number of fucose residues.</td>
</tr>
<tr>
<td>Gn Number of bisecting GlcNAc residues</td>
</tr>
<tr>
<td>Gal Number of galactose residues</td>
</tr>
<tr>
<td>Sia Number of sialic acid (NeuAc) residues</td>
</tr>
<tr>
<td>Br1 Extension level of branch 1.</td>
</tr>
<tr>
<td>Br2 Extension level of branch 2.</td>
</tr>
<tr>
<td>Br3 Extension level of branch 3.</td>
</tr>
<tr>
<td>Br4 Extension level of branch 4.</td>
</tr>
</tbody>
</table>

The full model generates 7565 N-glycan structures in a network of 22,871 reactions
### Distribution of carbohydrate chains in PDB
(Release September 2004)

<table>
<thead>
<tr>
<th>Chain length</th>
<th>N-glycan</th>
<th></th>
<th>O-glycans</th>
<th></th>
<th>ligands</th>
<th></th>
<th>total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1977</td>
<td>58.5</td>
<td>555</td>
<td>91.4</td>
<td>2093</td>
<td>59.0</td>
<td>4625</td>
<td>61.4</td>
</tr>
<tr>
<td>2</td>
<td>693</td>
<td>20.5</td>
<td>29</td>
<td>4.8</td>
<td>812</td>
<td>22.9</td>
<td>1534</td>
<td>20.4</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>9.2</td>
<td>10</td>
<td>1.7</td>
<td>329</td>
<td>9.3</td>
<td>649</td>
<td>8.6</td>
</tr>
<tr>
<td>4</td>
<td>83</td>
<td>2.5</td>
<td>2</td>
<td>0.3</td>
<td>149</td>
<td>4.2</td>
<td>234</td>
<td>3.1</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>2.9</td>
<td>5</td>
<td>0.8</td>
<td>83</td>
<td>2.3</td>
<td>186</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>98</td>
<td>2.9</td>
<td>2</td>
<td>0.3</td>
<td>42</td>
<td>1.2</td>
<td>142</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>55</td>
<td>1.7</td>
<td>4</td>
<td>0.7</td>
<td>15</td>
<td>0.4</td>
<td>74</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
<td>1.1</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0.4</td>
<td>52</td>
<td>0.7</td>
</tr>
<tr>
<td>total</td>
<td>3381</td>
<td>607</td>
<td>3550</td>
<td>7538</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“We need to be able to search databases for what is out there. Imagine genomics and proteomics without GenBank”

The current state of glyco-related databases can be characterized as “the biggest defect in the field”. (Ajit Varki).

**Recommendation 1:** Development of a robust, centralized, and thoroughly curated glycan structures database

To smooth the way for central carbohydrate structure database the active larger initiatives agreed to immediately start with the necessary preparatory steps for the conversion of CarbBank data into the GLYDE-II format
Summary

- Understanding protein modifications such as glycosylation is crucial to understand function
- Databases for Glyco-informatics Research is starting to come together
  - XML standardization
  - Major databases (Glycosciences.de, KEGG, CFG)
- More advanced informatics approaches can be applied to various facets of glyco-research
- Goal: to get the true overall picture of cellular processes
For further questions:

- Kiyoko F. Aoki-Kinoshiba
- kkiyoko@t.soka.ac.jp