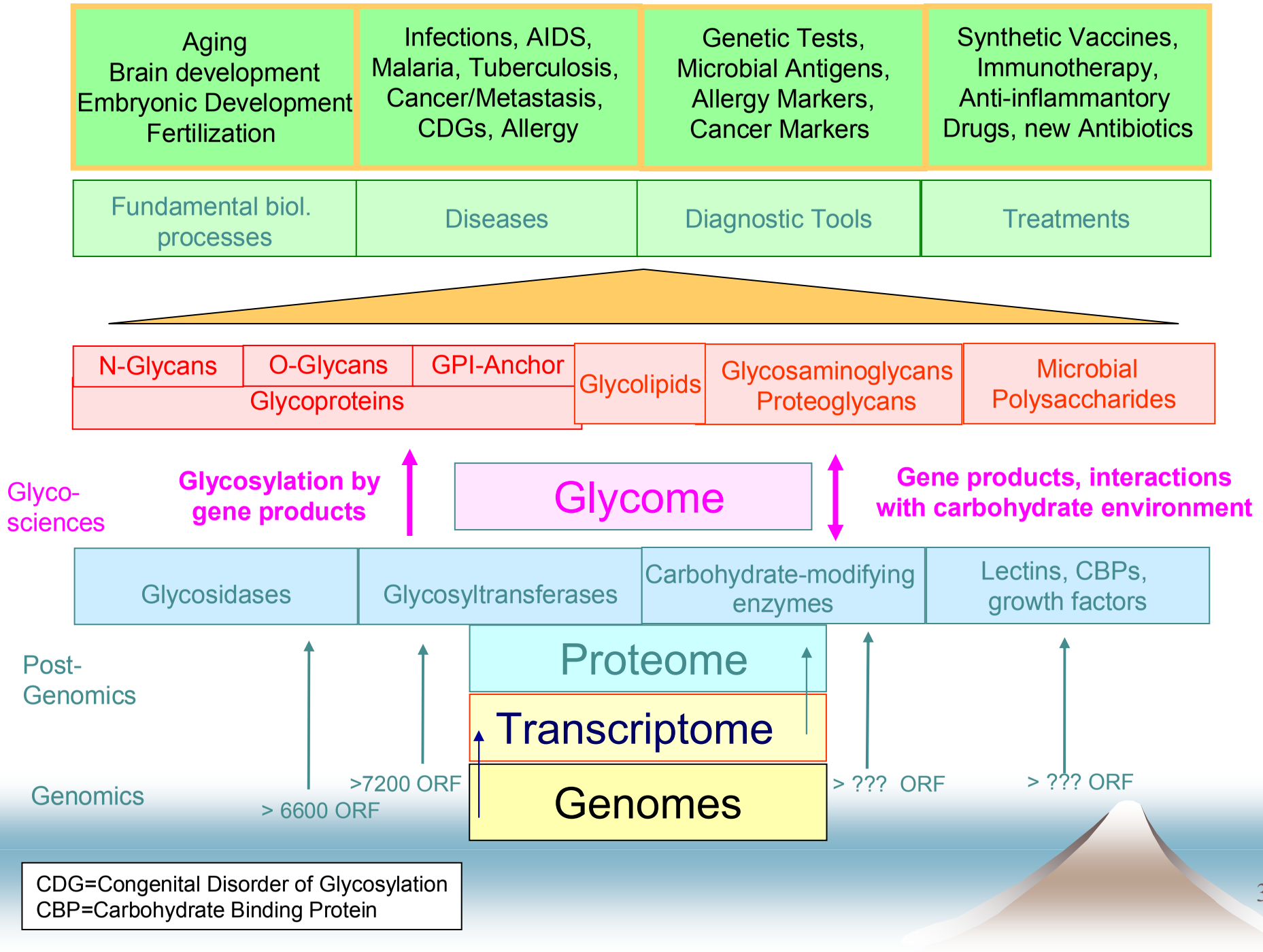
An introduction to bioinformatics for glycomics research

Kiyoko F. Aoki-Kinoshita

Soka University, Japan

Introduction

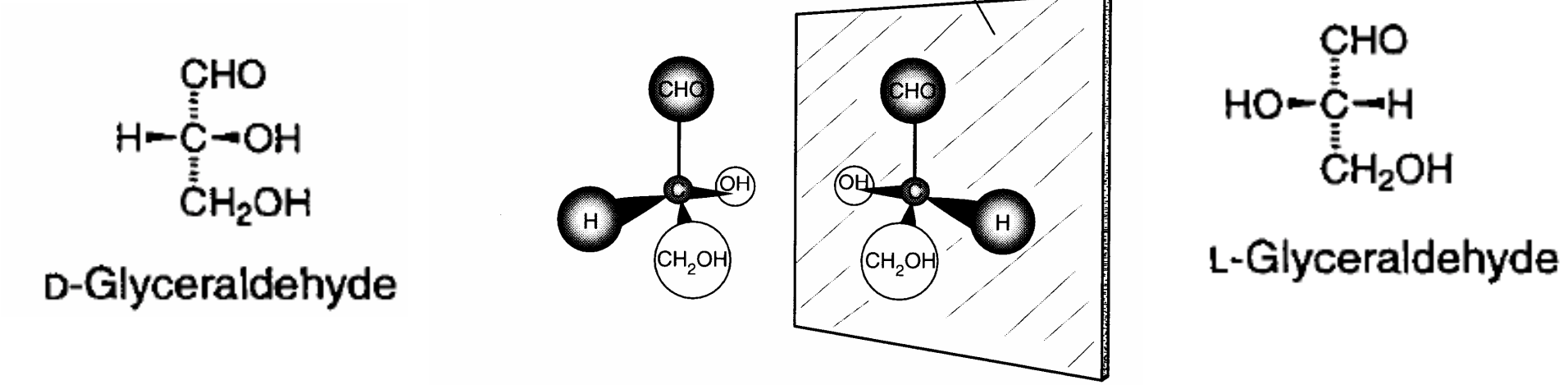
Biological role of carbohydrates as information containing molecules

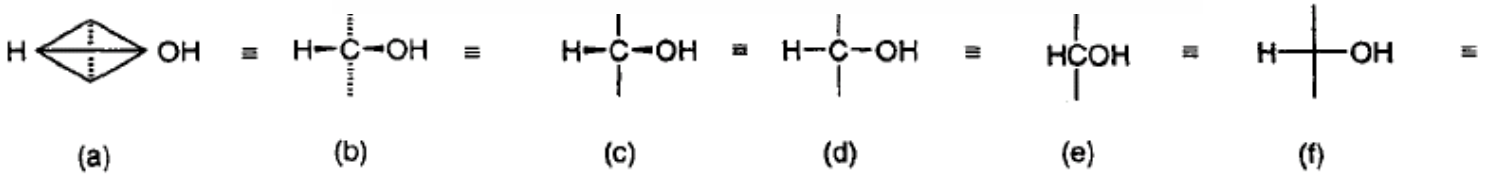

Aging
Brain development
Embryonic Development
Fertilization
Infections, AIDS,
Malaria, Tuberculosis,
Cancer/Metastasis,
CDGs, Allergy
Genetic Tests,
Microbial Antigens,
Allergy Markers,
Cancer Markers
Synthetic Vaccines,
Immunotherapy,
Anti-inflammantory
Drugs, new Antibiotics
Fundamental biol. processes
Diseases
Diagnostic Tools
Treatments
N-Glycans
O-Glycans
GPI-Anchor
Glycoproteins
Glycolipids
Glycosaminoglycans
Proteoglycans
Microbial
Polysaccharides
Glyco-
sciences
Glycosylation by
gene products
Glycome
Gene products, interactions
with carbohydrate environment
Glycosidases
Glycosyltransferases
Carbohydrate-modifying
enzymes
Lectins, CBPs,
growth factors
Post-
Genomics
Proteome
Transcriptome
Genomics
>7200 ORF
> 6600 ORF
Genomes
> ??? ORF
> ??? ORF
CDG=Congenital Disorder of Glycosylation
CBP=Carbohydrate Binding Protein

Introduction

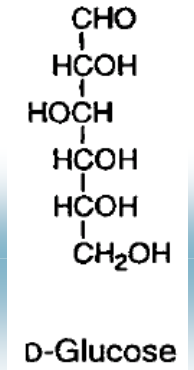
Nomenclature of Carbohydrates

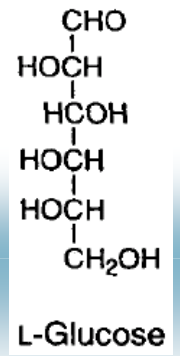
Glyceralaldehyde, the simplest Aldose, contains one chiral carbon atom carrying four different substituents and has therefore two different enantiomers.

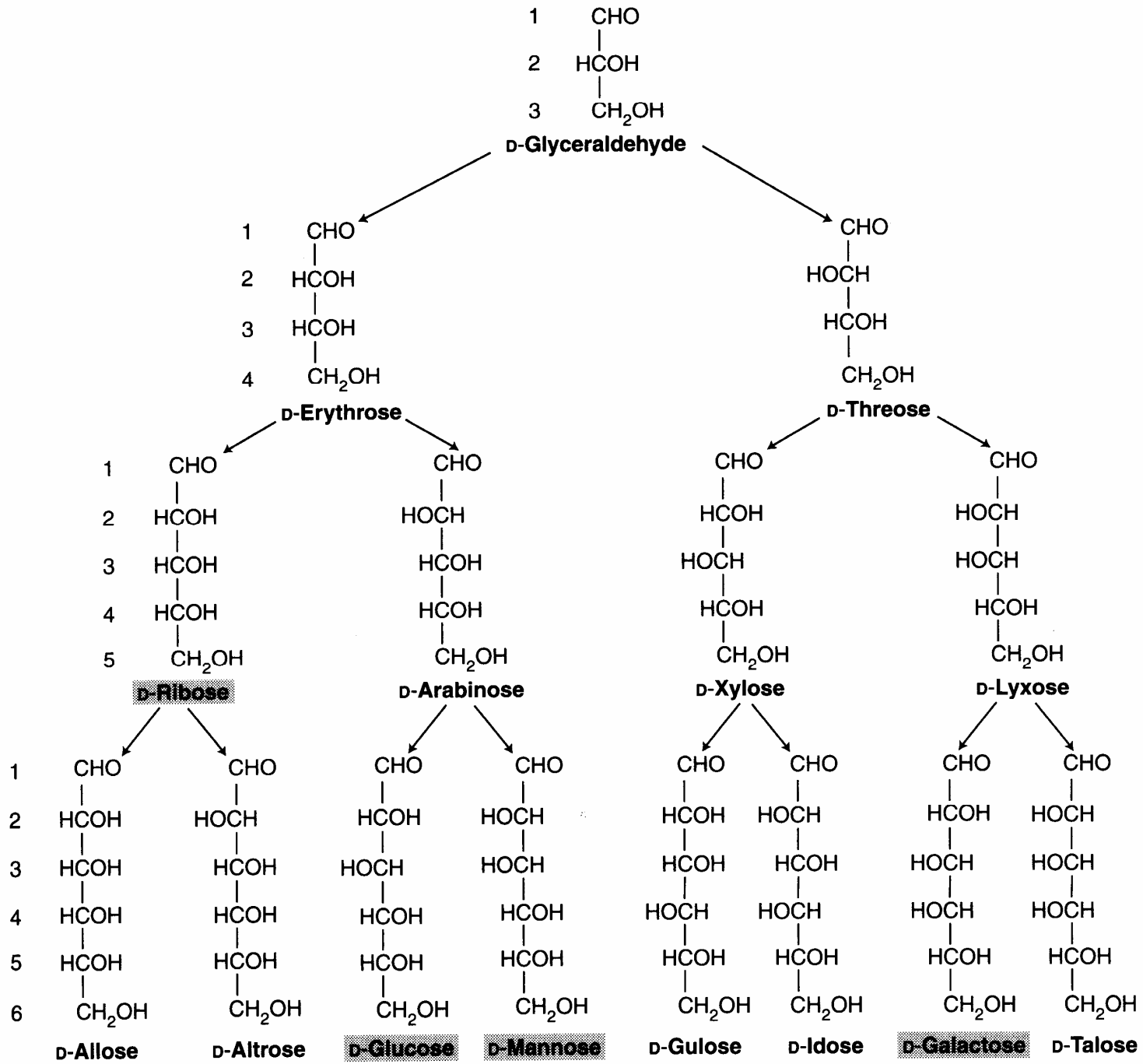




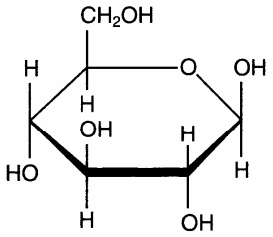
Conventional representation of a carbon atom (e.g. C-2 of D-glucose) in the Fischer projection.



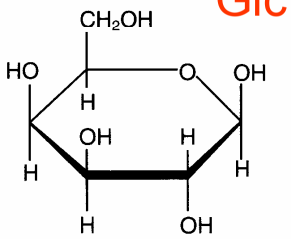



D-Glyceraldehyde
D-Erythrose
D-Threose
D-Ribose
D-Arabinose
D-Xylose
D-Lyxose
D-Allose
D-Altrose
D-Glucose
D-Mannose
D-Gulose
D-Idose
D-Galactose
D-Talose

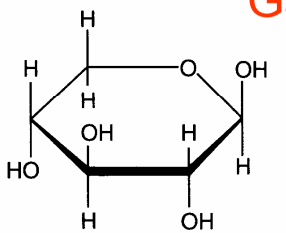
Some common and biologically important monosaccharides


β-D-Glucose

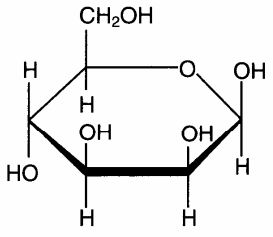
Glc


β-D-Galactose

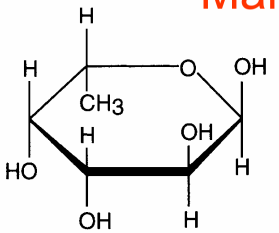
Gal


β-D-Xylose

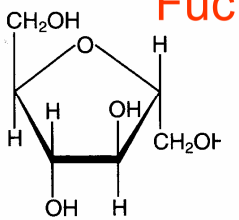
Xyl


β-D-Mannose

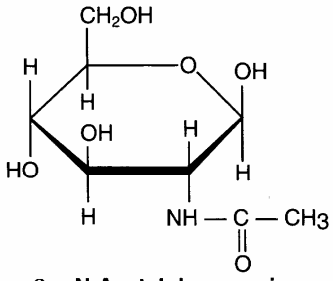
Man


β-L-Fucose

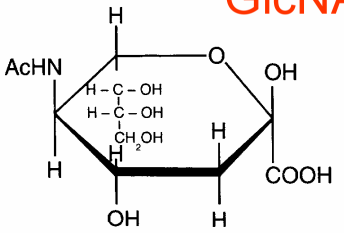
Fuc


β-D-Frutose

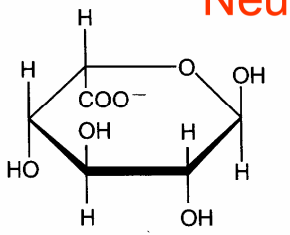
Fru


β-D-N-Acetylglucosamine

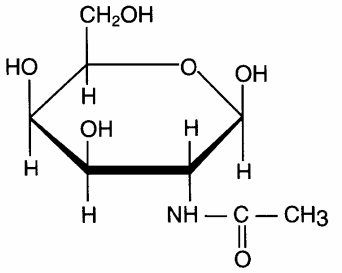
GlcNAc


β-N-Acetyl neuraminic acid

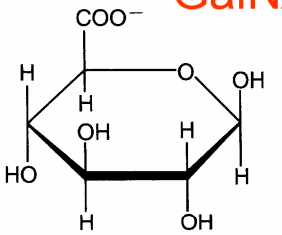
Neu5NAc


L-β-Iduronic acid

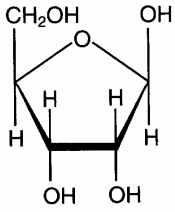
IdoA


β-D-N-Acetylgalactosamine

GalNAc


β-D-Glucuronic acid

GalA


β-D-Ribose

Rib

Reference Database of Monosaccharides
http://www.monosaccharidedb.org

Exact query:

MonoSaccharideDB
home | notation | query
• exact search • fuzzy search

Enter a residue name following the **rules for monosaccharide residue notation**:

Submit Query | Reset

Or build residue name from pull down menus:

Anomeric: a (alpha) (This field

Abs. Config: L

Base Type: Fuc (Fucose)

Ring Type: p (pyranose)

Uronic Acid: Tick checkbox to en

Modifications: (For a definition of the n

Position: Type: deoxy (explic

Position: 3 Type: fluoro (Fluori

Position: Type: deoxy (explic

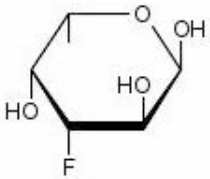
Position: Type: deoxy (explic

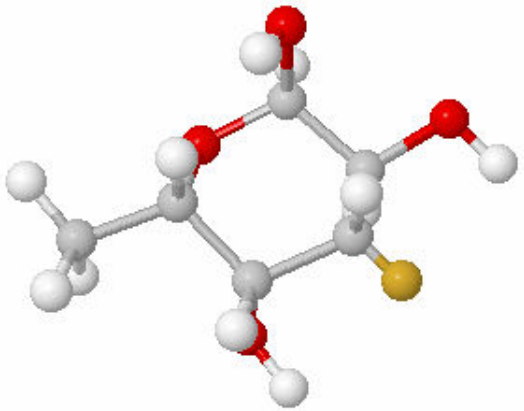
Submit Query | Reset

Monosaccharide: a-L-Fucp3fluoro

Properties:	
ID:	228
Name:	a-L-Fucp3fluoro
BaseType:	Fuc (Fucose)
RootType:	Gal
Size:	6
Anomeric:	a
Abs. Config:	L
Ring Type:	p
Stereocode:	22112

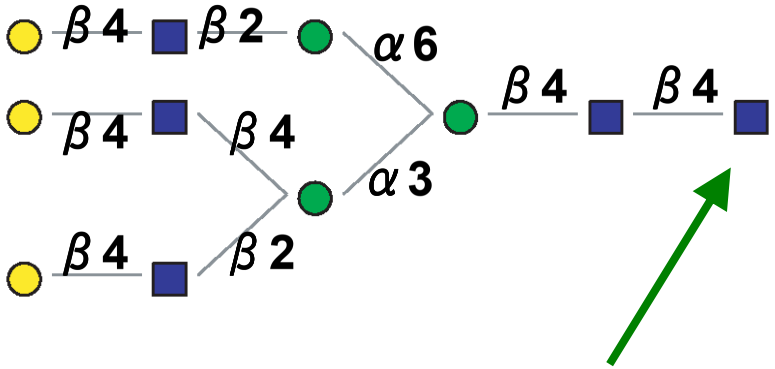
Modifications:	
Name:	fluoro (Fluorine atom)
Position:	3





Oligosaccharide description

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




```
                                a-D-Manp-(1-6)+
                                      |       |
b-D-Galp-(1-4)-b-D-GlcpNAc-(1-2)+     |       |
                                      |       |
                                    b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
b-D-Galp-(1-4)-b-D-GlcpNAc-(1-2)+     |       |
                                      |       |
                                a-D-Manp-(1-3)+
                                      |
b-D-Galp-(1-4)-b-D-GlcpNAc-(1-4)+
```

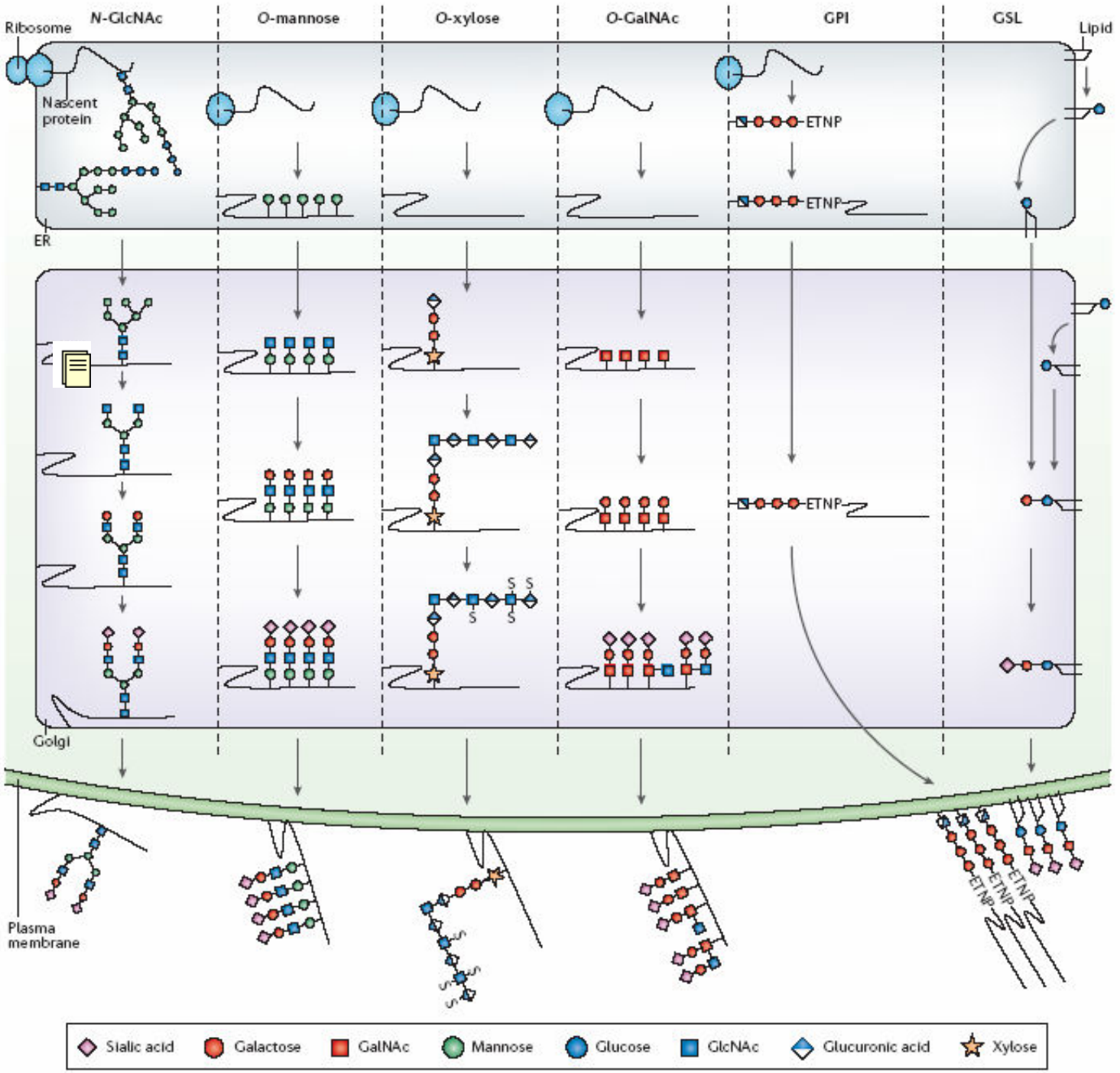
Introduction

Glyco-related pathways

Overview of glycan biosynthetic pathways.

Hudson H Freeze
Genetic defects in the human glycome.
Nat Rev Genet.

2006 Jul;7(7):537-51



Glycan classes: functions and biosynthesis

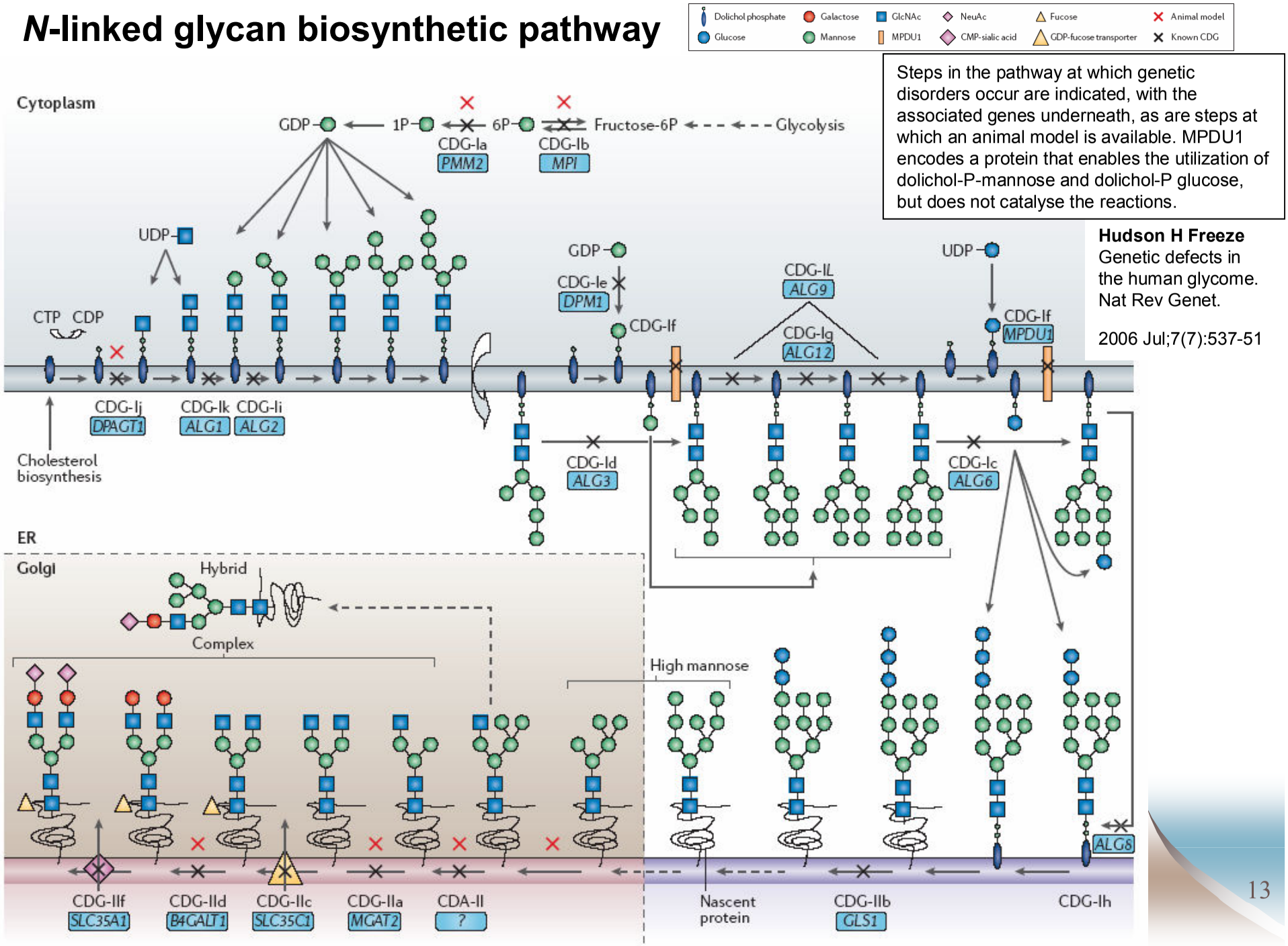
Glycan Type	Linkage	Synthesis	Functions
Protein			
N-Linked	GlcNAc-β-asparagine	Added in ER, modified in Golgi	Protein folding and stability; complex formation; signalling; cell–cell recognition
O-Linked	Mannose-α-serine/threonine	Begins in ER, completed in Golgi	Stability and function of dystrophin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

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.

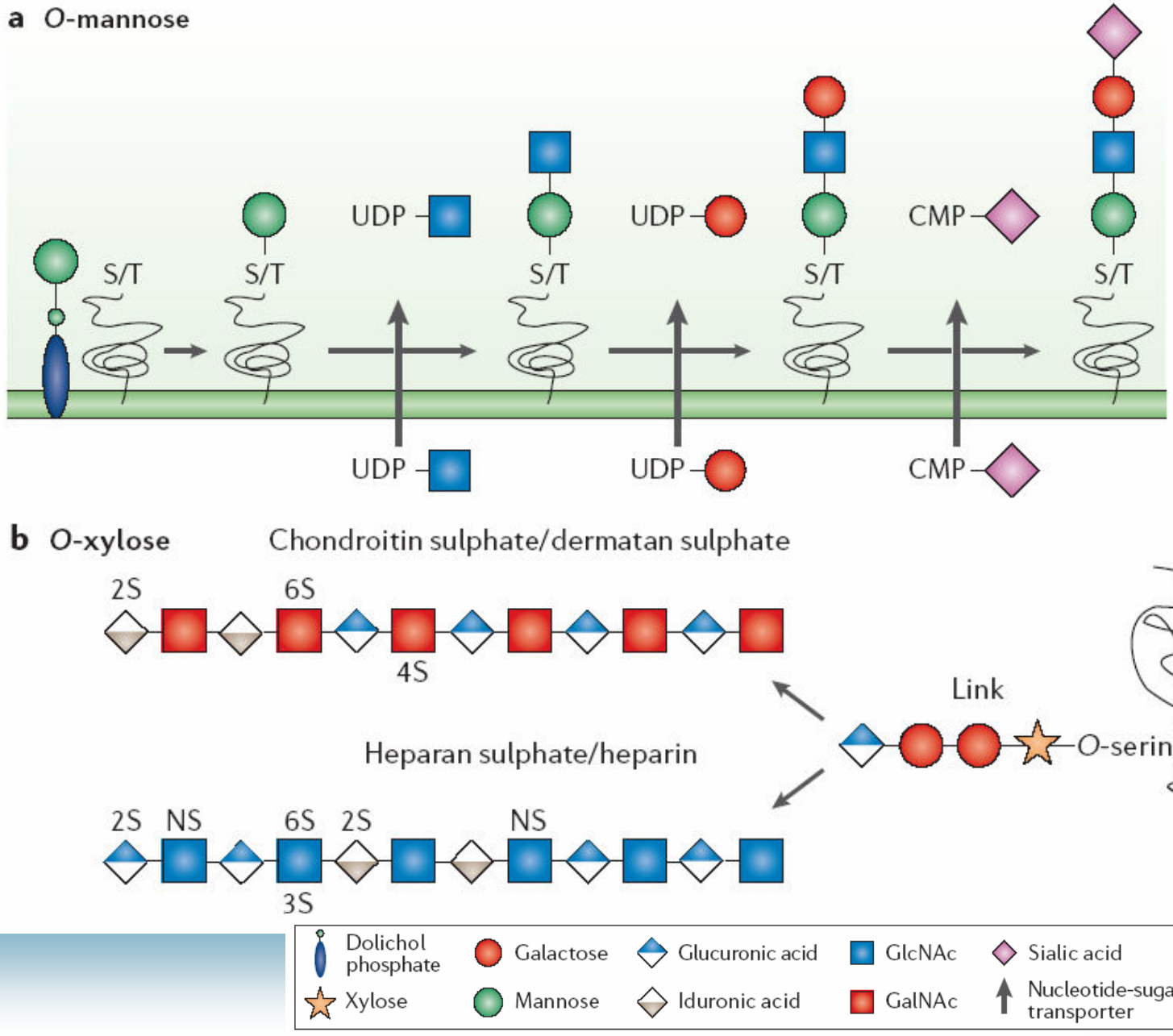
Hudson H Freeze
Genetic defects in the human glycome.
Nat Rev Genet.

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

O-mannose and *O*-xylose biosynthetic pathways



NS, 2S, 3S, 4S and 6S represent 2-N-, 2-O-, 3-O-, 4-O- and 6-O-sulphate, in that order.

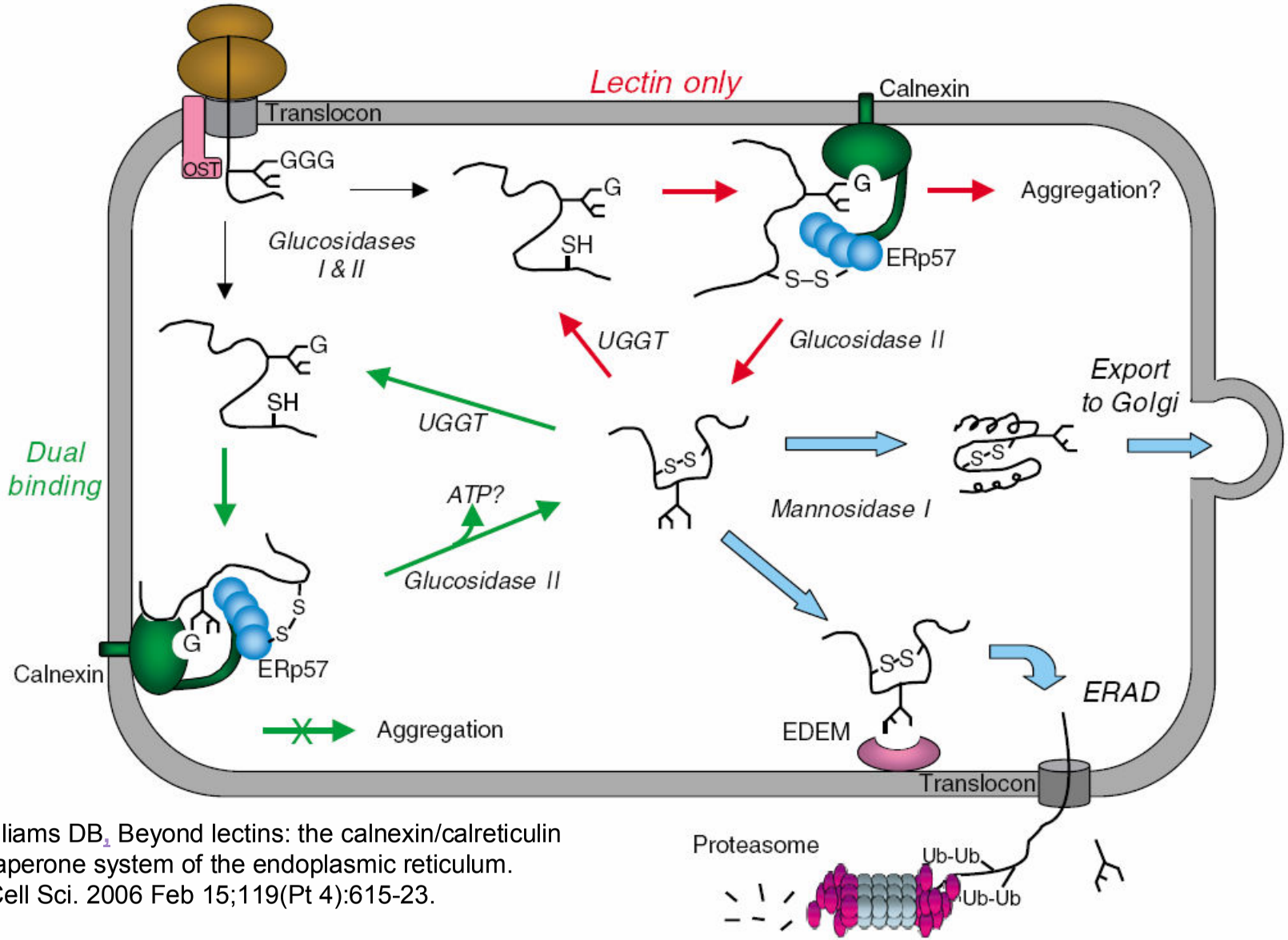
Human diseases caused by genetic defects in O-glycosylation pathways

- 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

- Paroxysomal 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.



Williams DB. Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum. J Cell Sci. 2006 Feb 15;119(Pt 4):615-23.

Glyco-databases and data formats

Carbohydrate Structure Databases

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

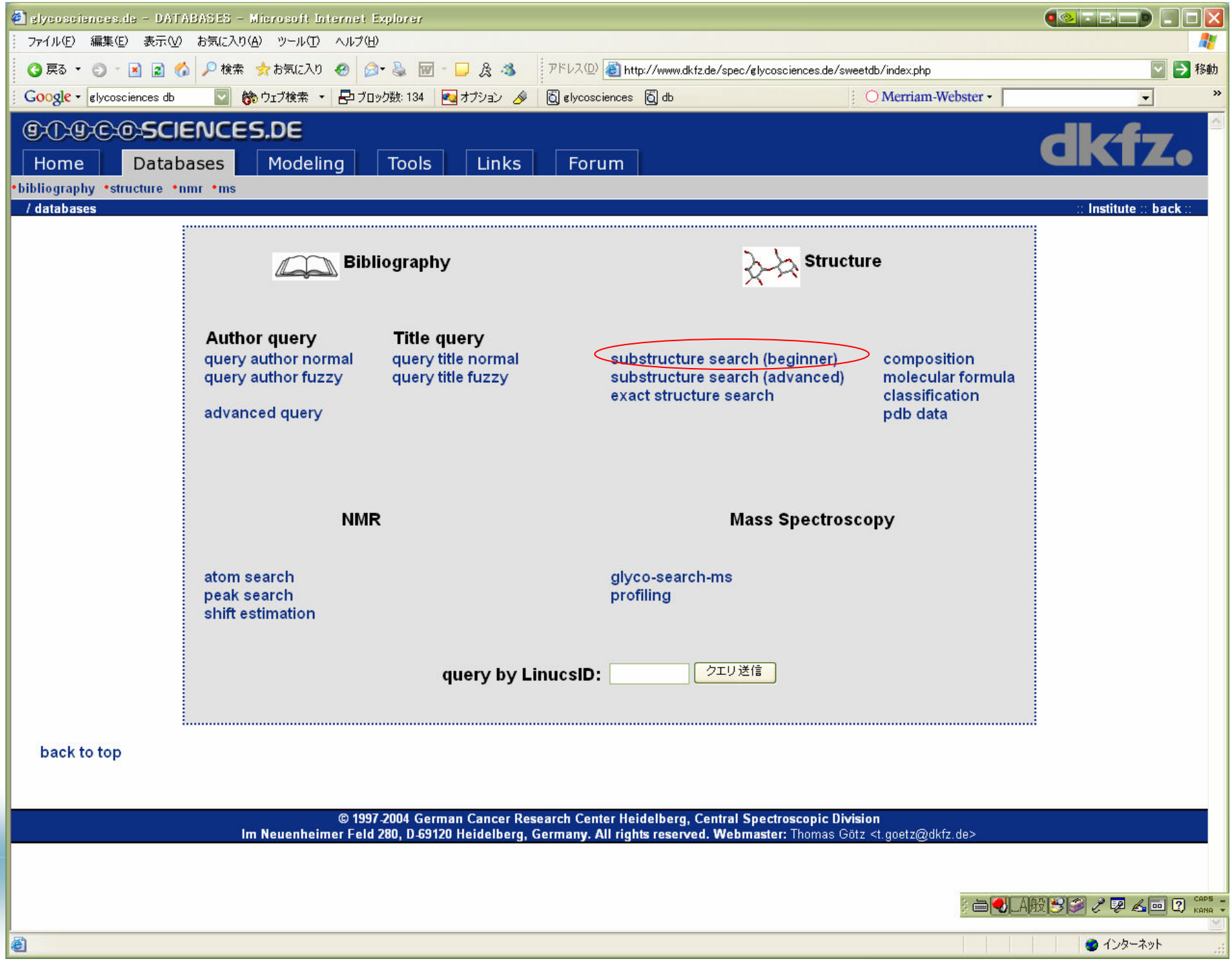
- Commercial databases:
  - GlycoSuite (Proteome Systems, Ltd.)
  - Glycomics DB (Glycominds, Ltd.)

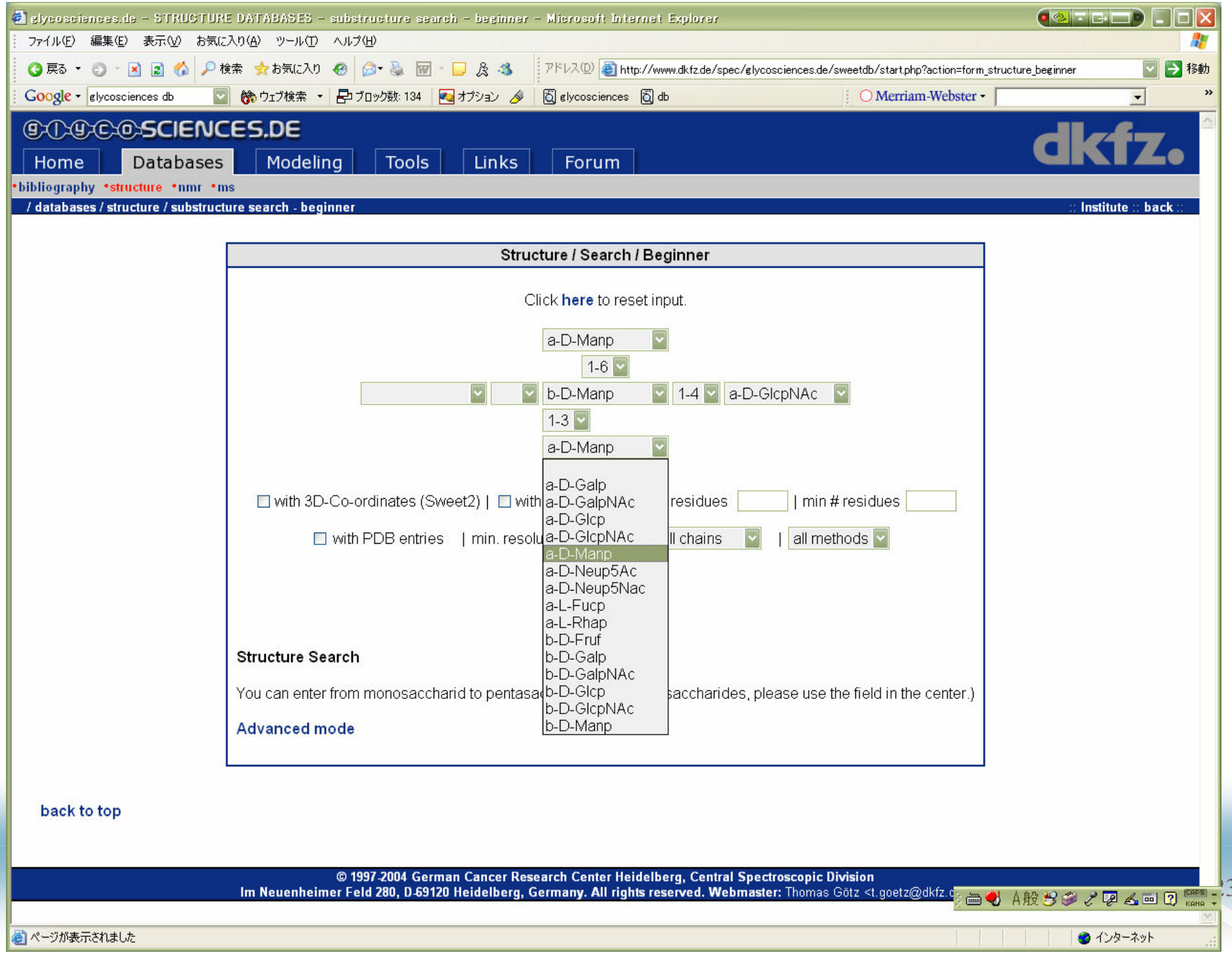
CarbBank

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

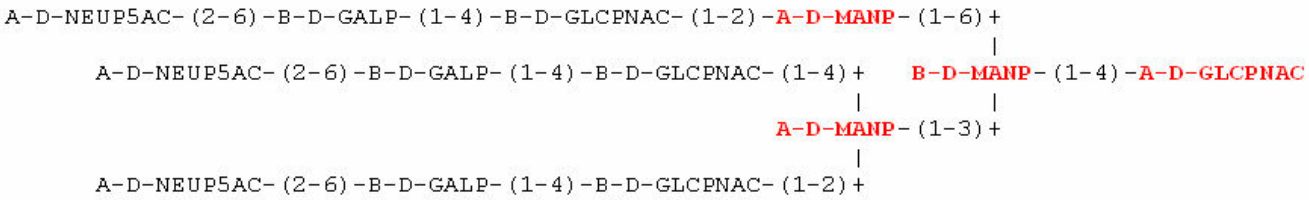
GLYCOSCIENCES.de DB

- 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

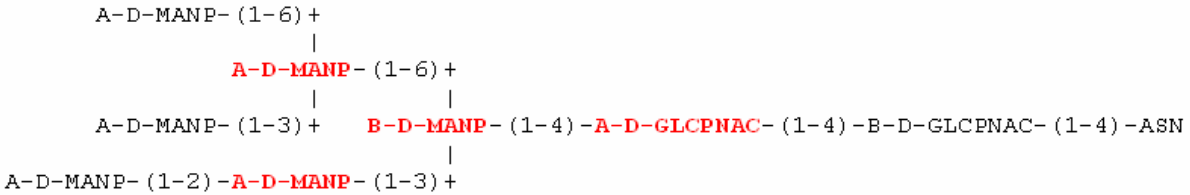

glycosciences.de - DATABASES - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
http://www.dkfz.de/spec/glycosciences.de/sweetdb/index.php
glycosciences db
glycosciences
db
Merriam-Webster
GLYCOSCIENCES.DE
dkfz.
Home
Databases
Modeling
Tools
Links
Forum
•bibliography •structure •nmr •ms
/ databases
:: Institute :: back ::
Bibliography
Structure
Author query
query author normal
query author fuzzy
advanced query
Title query
query title normal
query title fuzzy
substructure search (beginner)
substructure search (advanced)
exact structure search
composition
molecular formula
classification
pdb data
NMR
atom search
peak search
shift estimation
Mass Spectroscopy
glyco-search-ms
profiling
query by LinucsID:
クエリ送信
back to top
© 1997-2004 German Cancer Research Center Heidelberg, Central Spectroscopic Division
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany. All rights reserved. Webmaster: Thomas Götz <t.goetz@dkfz.de>
インターネット


glycosciences.de - STRUCTURE DATABASES - substructure search - beginner - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
http://www.dkfz.de/spec/glycosciences.de/sweetdb/start.php?action=form_structure_beginner
GLYCOSCIENCES.DE
dkfz.
Home
Databases
Modeling
Tools
Links
Forum
•bibliography •structure •nmr •ms
/ databases / structure / substructure search - beginner
:: Institute :: back ::
Structure / Search / Beginner
Click here to reset input.
a-D-Manp
1-6
b-D-Manp
1-4
a-D-GlcpNAc
1-3
a-D-Manp
a-D-Galp
a-D-GalpNAc
a-D-Glcp
a-D-GlcpNAc
a-D-Manp
a-D-Neup5Ac
a-D-Neup5Nac
a-L-Fucp
a-L-Rhap
b-D-Fruf
b-D-Galp
b-D-GalpNAc
b-D-Glcp
b-D-GlcpNAc
b-D-Manp
with 3D-Co-ordinates (Sweet2)
with PDB entries
min. resolu
residues
min # residues
ll chains
all methods
Structure Search
You can enter from monosaccharid to pentasa
saccharides, please use the field in the center.)
Advanced mode
back to top
© 1997-2004 German Cancer Research Center Heidelberg, Central Spectroscopic Division
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany. All rights reserved. Webmaster: Thomas Götz
ページが表示されました
インターネット

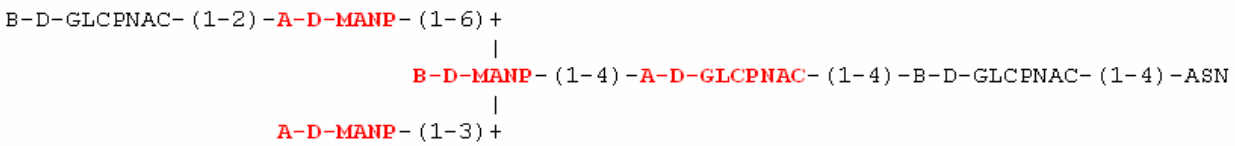
Searched structures for substructures. Results: 1 - 10 of 13



Explore



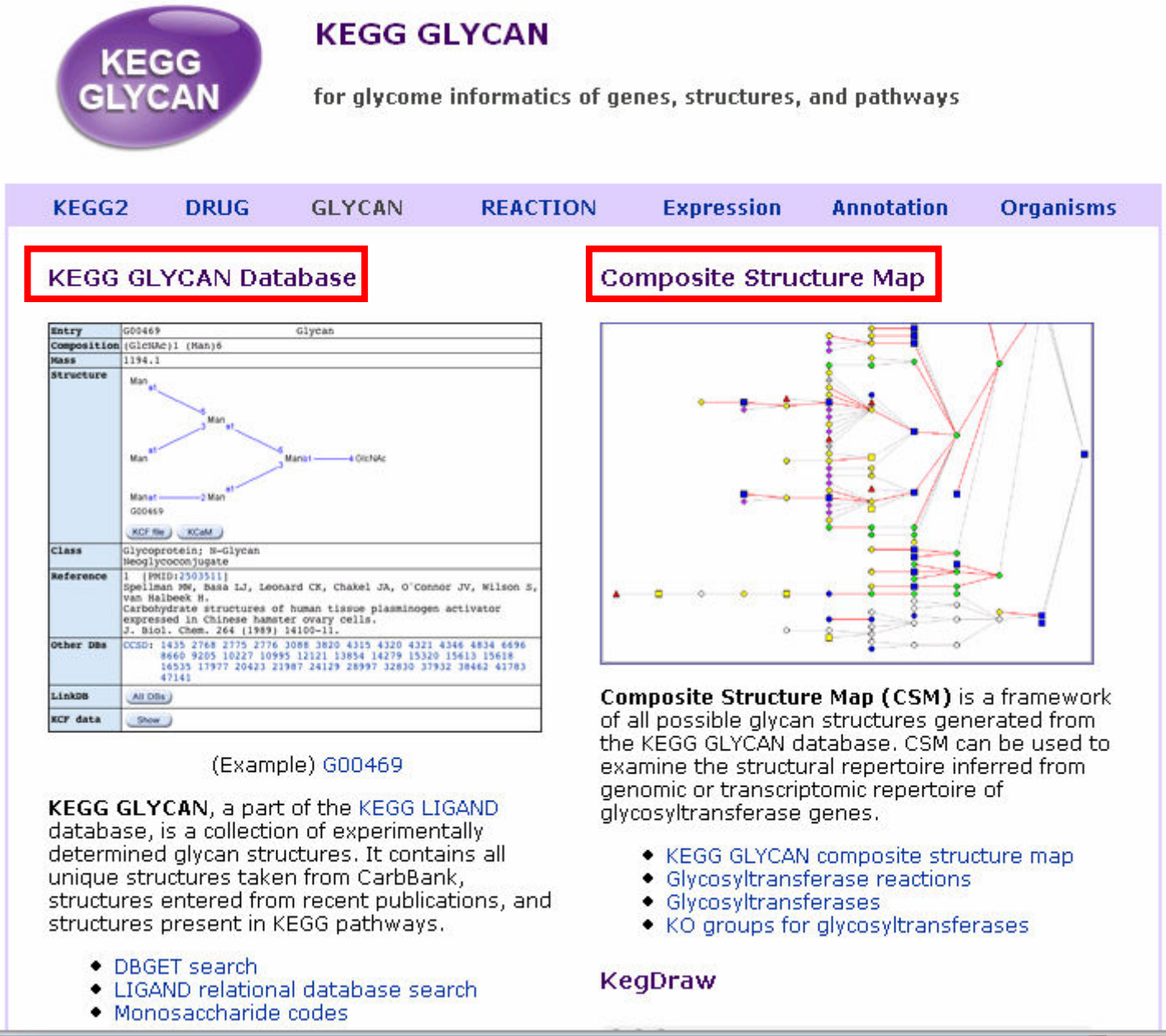
Explore pdb-entries



Explore pdb-entries

KEGG GLYCAN

- http://www.genome.jp/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



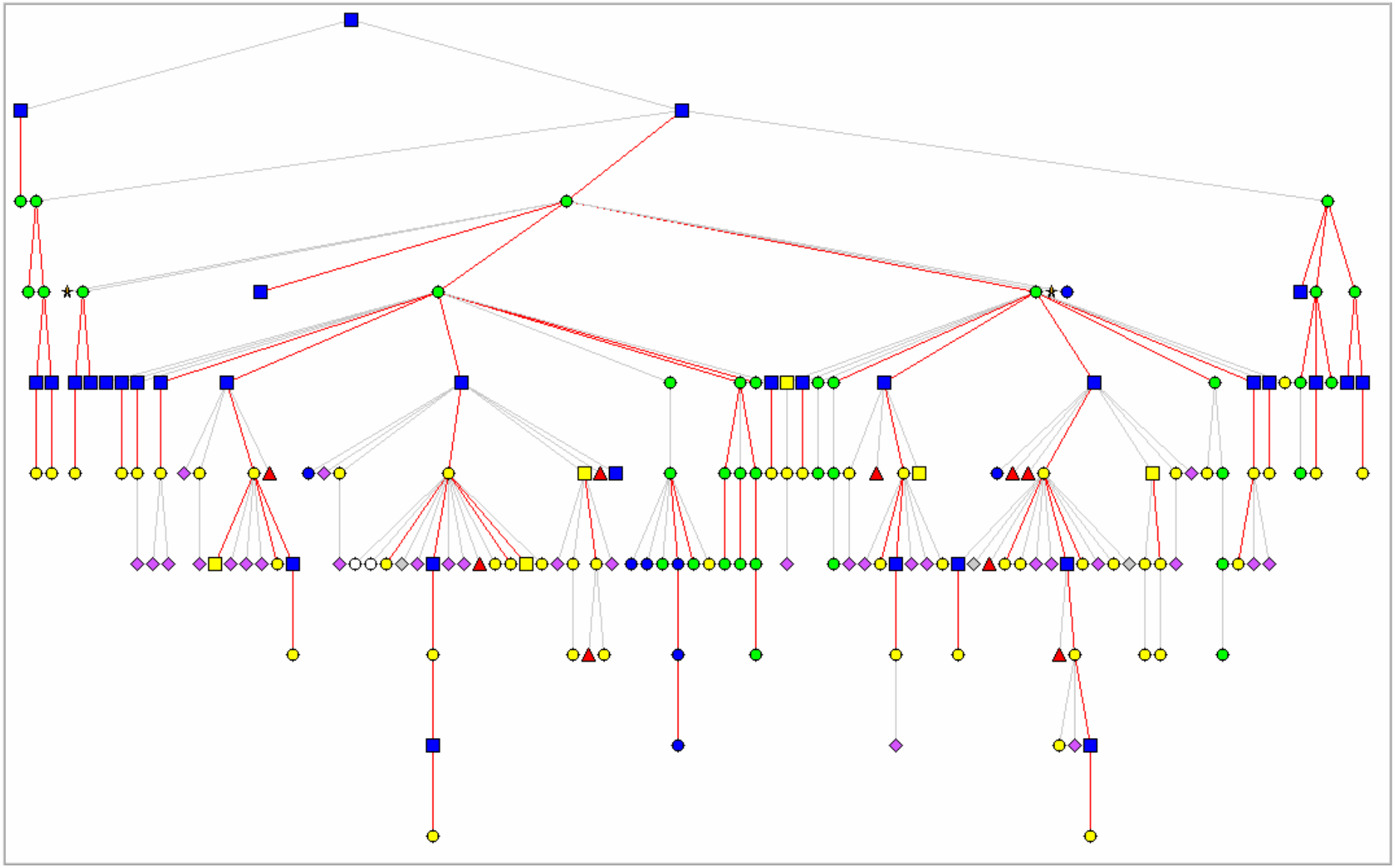
- http://www.genome.jp/kegg/glycan /

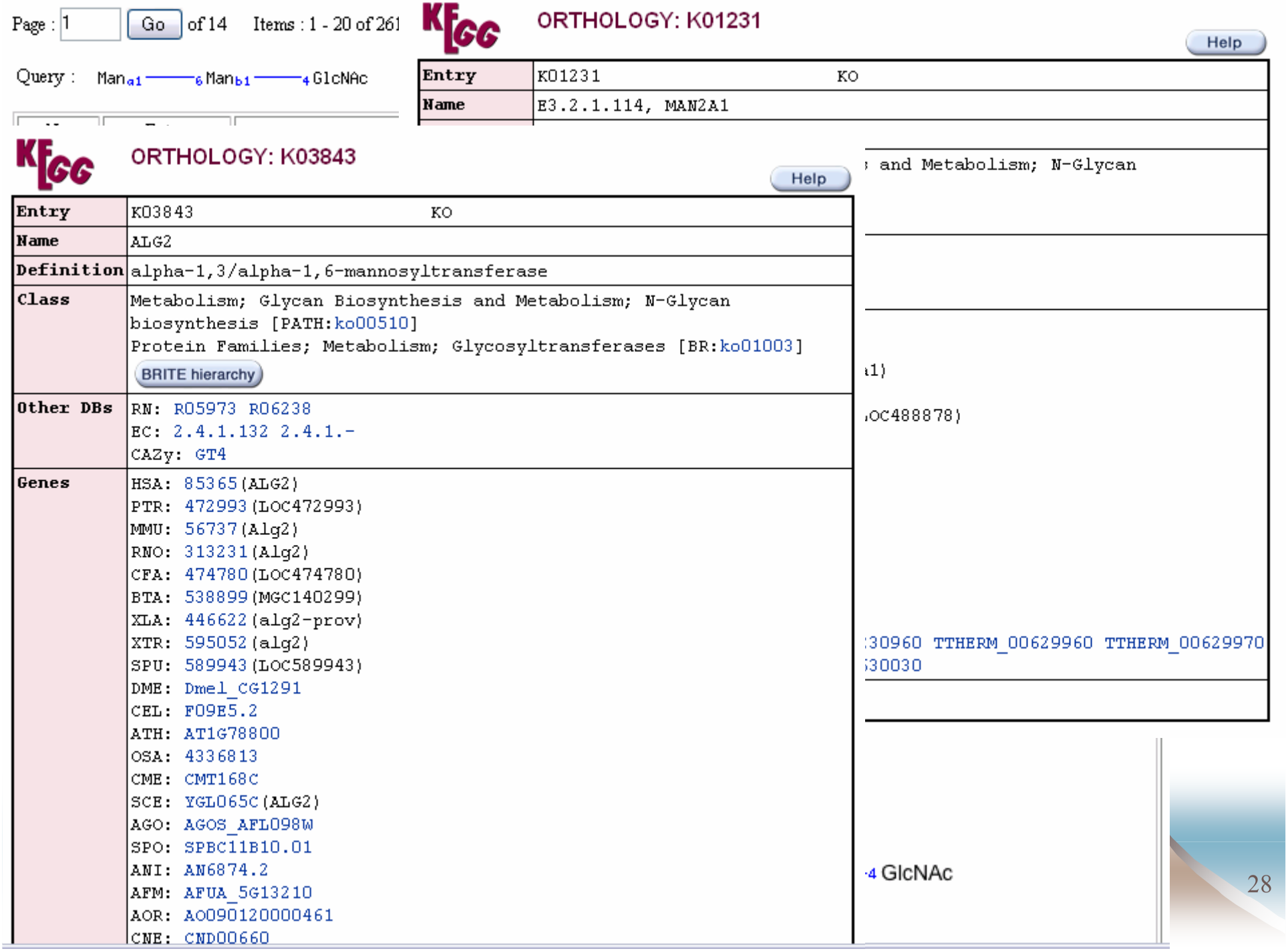
KEGG Glycan Composite Structure Map

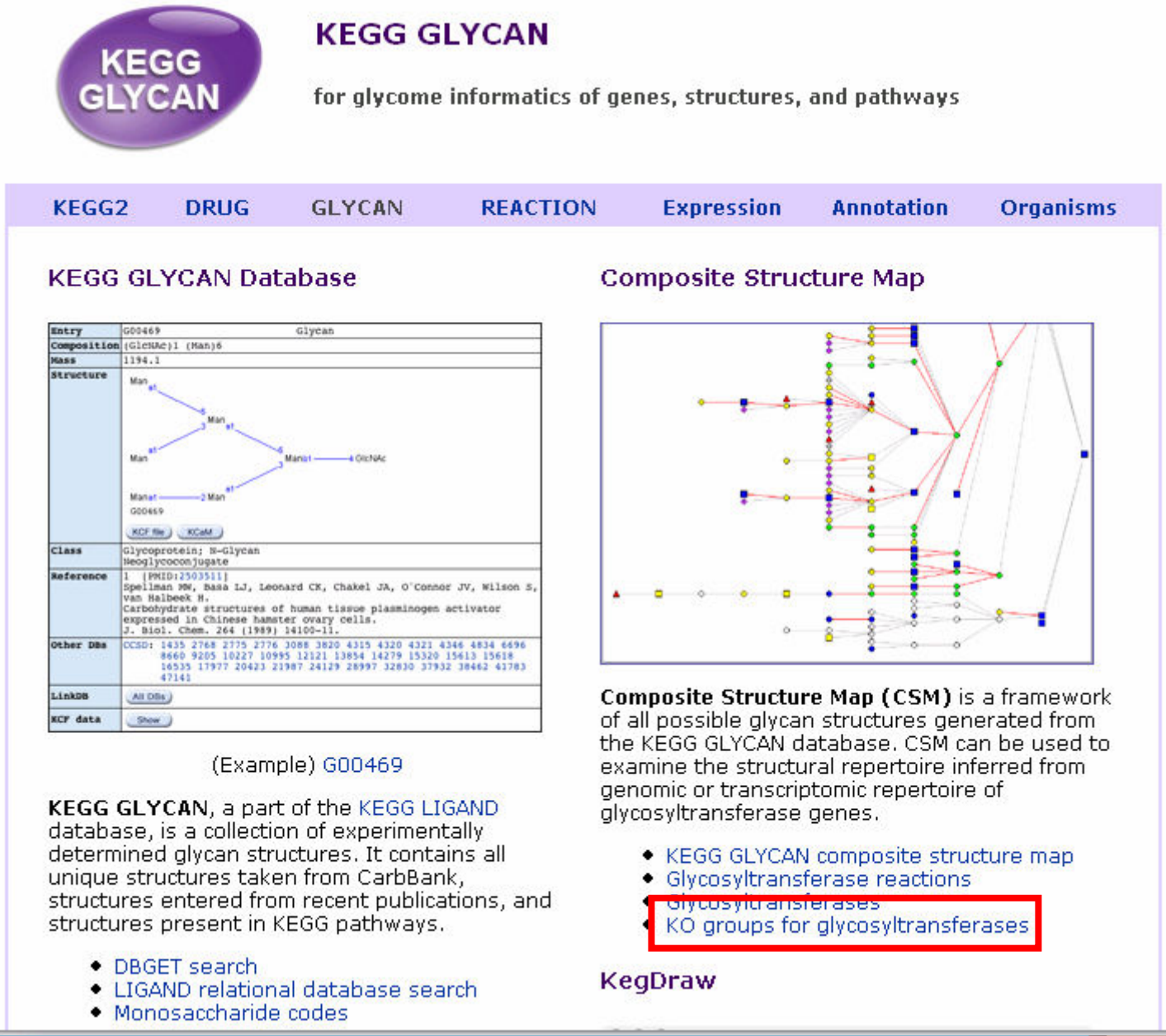
Help

Next2: Man - Next1: GlcNAc - Root: GlcNAc Threshold: 3 Position: Top Node size: 5

Homo sapiens Go Reset




Page : 1 Go of 14 Items : 1 - 20 of 261
Query : Man a1 6 Man b1 4 GlcNAc
KEGG ORTHOLOGY: K01231
Help
Entry K01231 KO
Name E3.2.1.114, MAN2A1
KEGG ORTHOLOGY: K03843
Help
Entry K03843 KO
Name ALG2
Definition alpha-1,3/alpha-1,6-mannosyltransferase
Class Metabolism; Glycan Biosynthesis and Metabolism; N-Glycan biosynthesis [PATH:ko00510]
Protein Families; Metabolism; Glycosyltransferases [BR:ko01003]
BRITE hierarchy
Other DBs RN: R05973 R06238
EC: 2.4.1.132 2.4.1.-
CAZy: GT4
Genes HSA: 85365(ALG2)
PTR: 472993(LOC472993)
MMU: 56737(Alg2)
RNO: 313231(Alg2)
CFA: 474780(LOC474780)
BTA: 538899(MGC140299)
XLA: 446622(alg2-prov)
XTR: 595052(alg2)
SPU: 589943(LOC589943)
DME: Dmel_CG1291
CEL: F09E5.2
ATH: AT1G78800
OSA: 4336813
CME: CMT168C
SCE: YGL065C(ALG2)
AGO: AGOS_AFL098W
SPO: SPBC11B10.01
ANI: AN6874.2
AFM: AFUA_5G13210
AOR: AO090120000461
CNE: CND00660
and Metabolism; N-Glycan
a1)
LOC488878)
30960 TTHERM_00629960 TTHERM_00629970
30030
4 GlcNAc



- http://www.genome.jp/kegg/glycan /

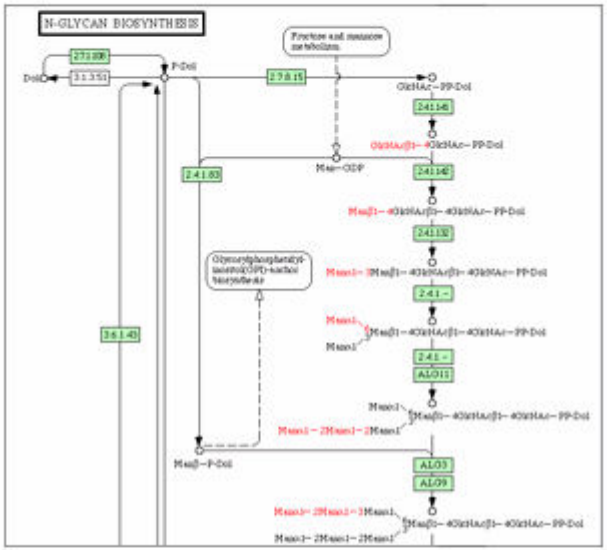
KO (KEGG Orthology) groups for glycosyltransferases

N-glycan biosynthesis

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

- DBGET search
- LIGAND relational database search
- Monosaccharide codes

KEGG GLYCAN Pathway Map



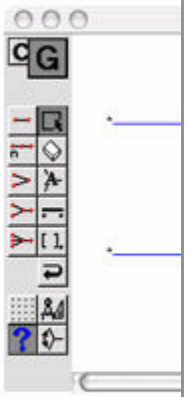
(Example) sce00510

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

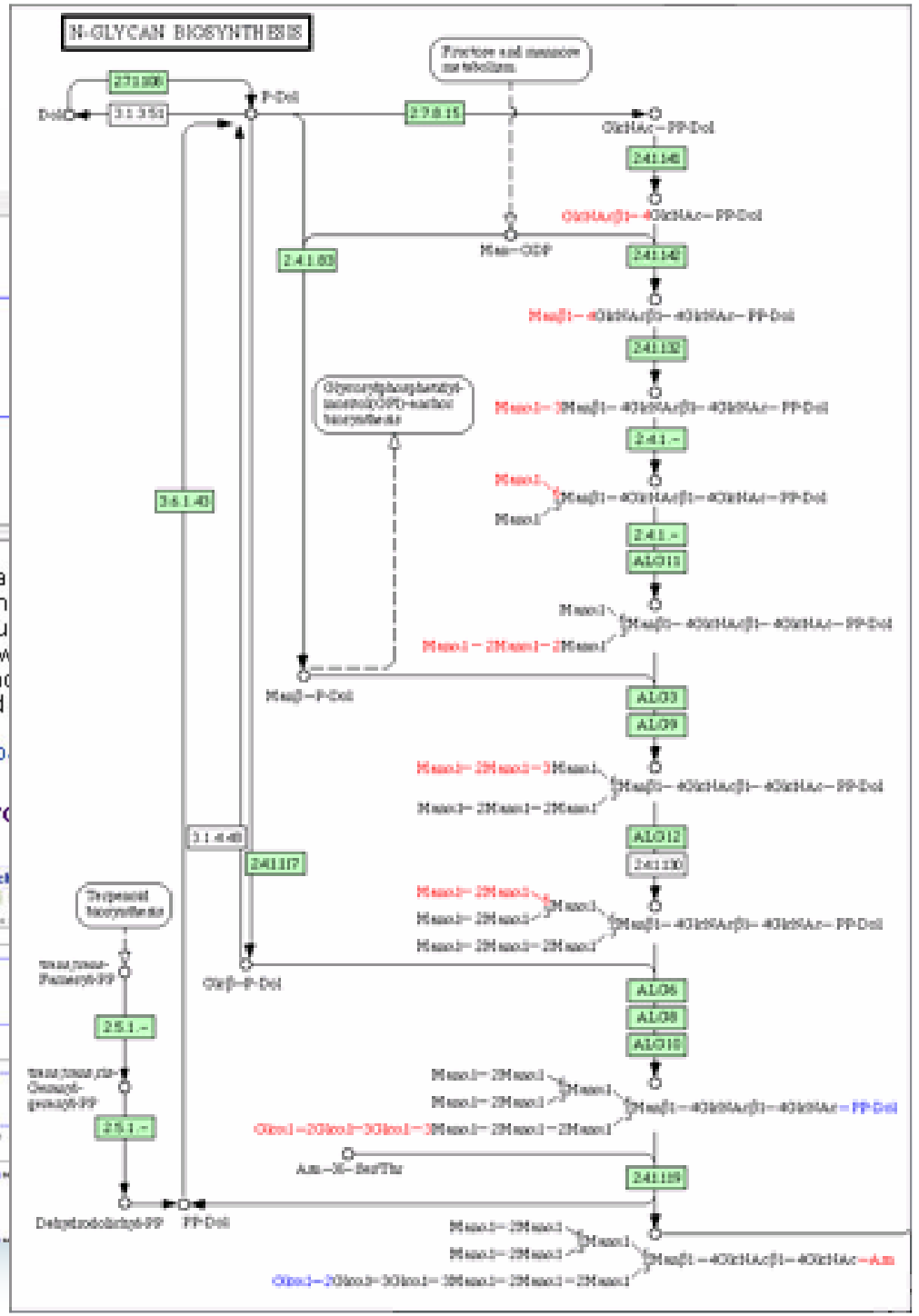
KegDraw



KegDraw is a
drawing chem
glycan structu
and ISIS/Draw
and Linux, and
academic and

- Downlo

KCaM Searc

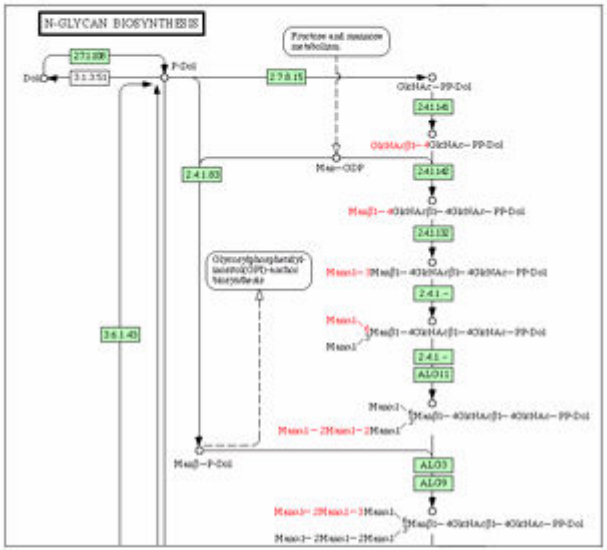


KEGG's Glycan Biosynthesis and Metabolism Pathways

N-Glycan biosynthesis
High-mannose type N-glycan biosynthesis
N-Glycan degradation
O-Glycan biosynthesis
Chondroitin / heparan sulfate biosynthesis
Keratan sulfate biosynthesis
Glycosaminoglycan degradation
Lipopolysaccharide biosynthesis
Peptidoglycan biosynthesis
Glycosylphosphatidylinositol(GPI) -anchor biosynthesis
Glycosphingolipid metabolism
Blood group glycolipid biosynthesis - lactoseries
Blood group glycolipid biosynthesis - neo-lactoseries
Globoside metabolism
Ganglioside biosynthesis
Glycan structures - biosynthesis 1
Glycan structures - biosynthesis 2
Glycan structures - degradation

- DBGET search
- LIGAND relational database search
- Monosaccharide codes

KEGG GLYCAN Pathway Map



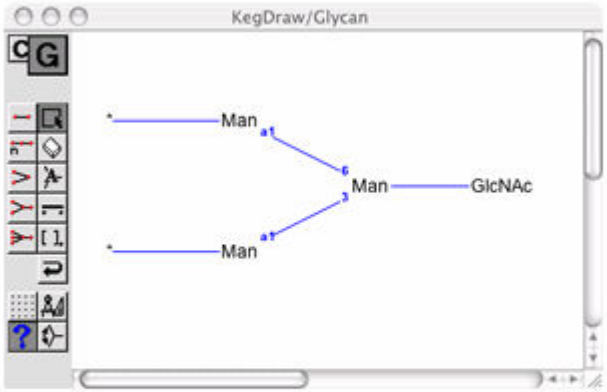
(Example) sce00510

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

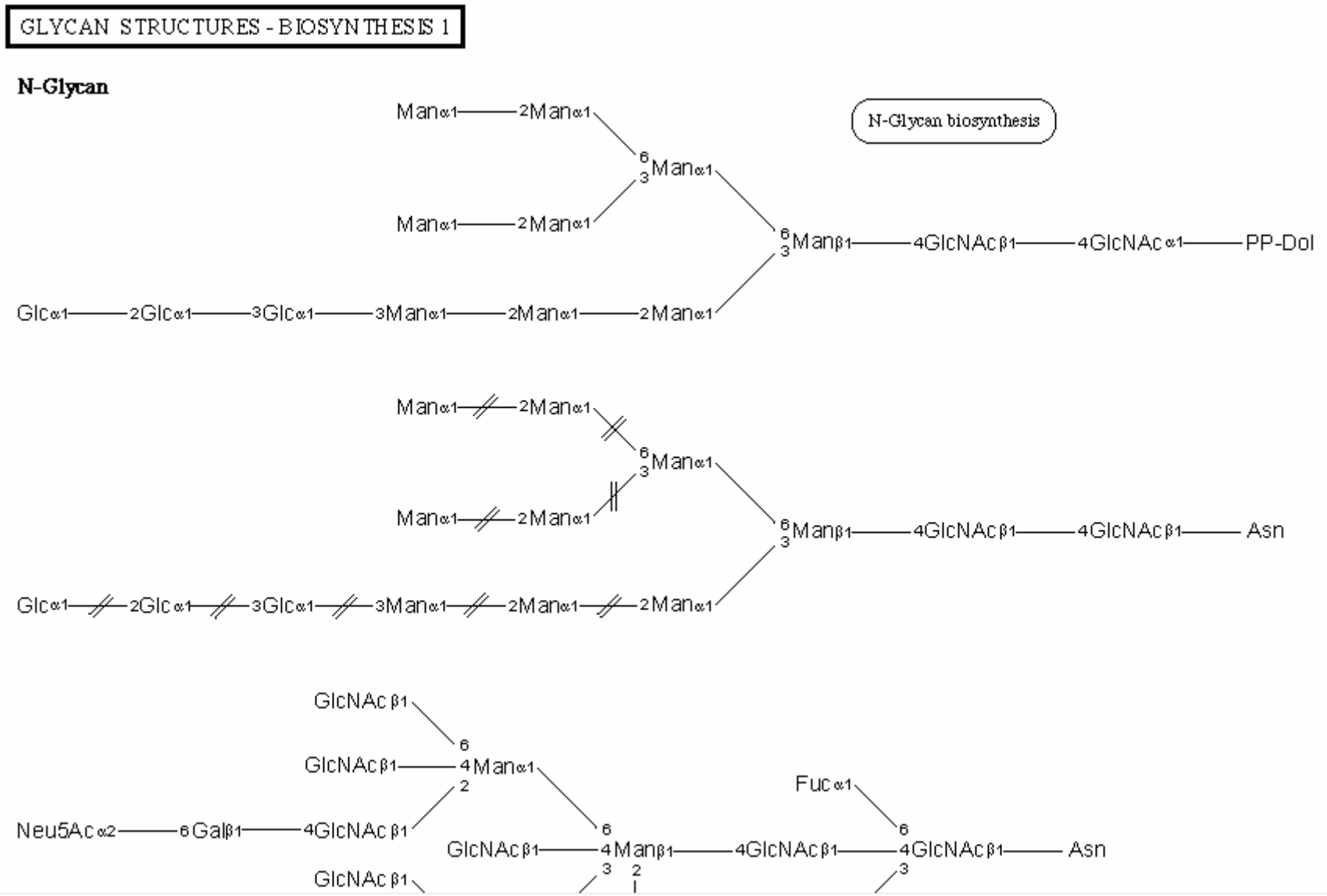


KegDraw is a standalone Java application for drawing chemical compound structures and glycan structures in a way similar to ChemDraw and ISIS/Draw. KegDraw runs on Mac, Windows, and Linux, and is made freely available to both academic and non-academic users.

- Download KegDraw

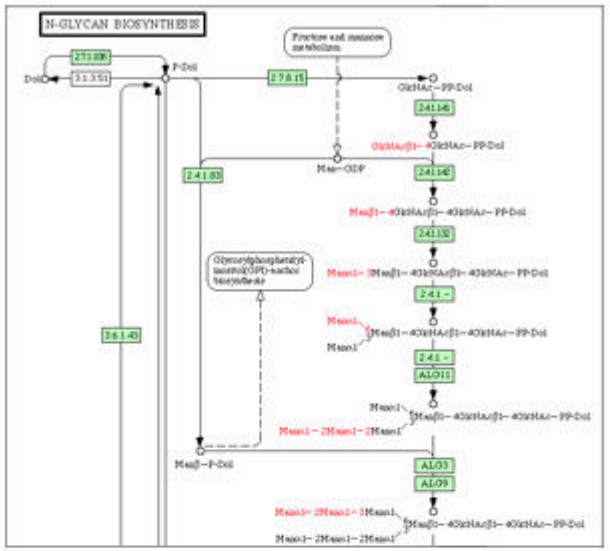
KCaM Search

KEGG Glycan Structure Map



- DBGET search
- LIGAND relational database search
- Monosaccharide codes

KEGG GLYCAN Pathway Map



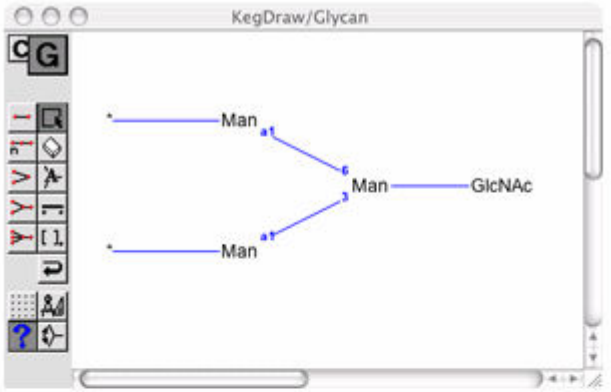
(Example) sce00510

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

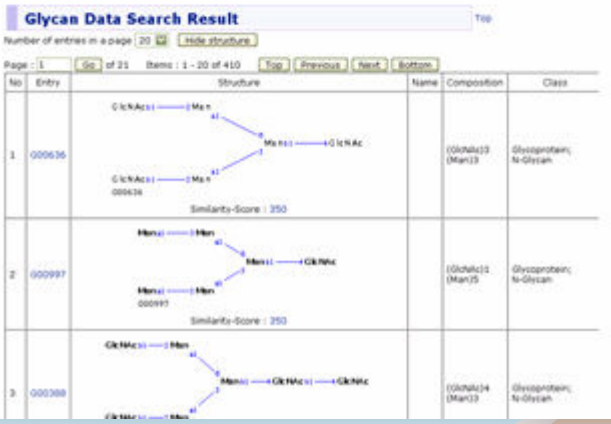
KegDraw



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

Compute Clear

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

Glycan ID **G00078** (Example) G00021 View structure

KCF File Name 参照...

KCF File Text

KCaM Main Server
KCaM Tutorial
KCaM FAQ
KCaM Docs

Select target database:

KEGG GLYCAN CarbBank

Select program:

Gapped (Approximate match)
Ungapped (Exact match)

Select option:

Global search
Local search

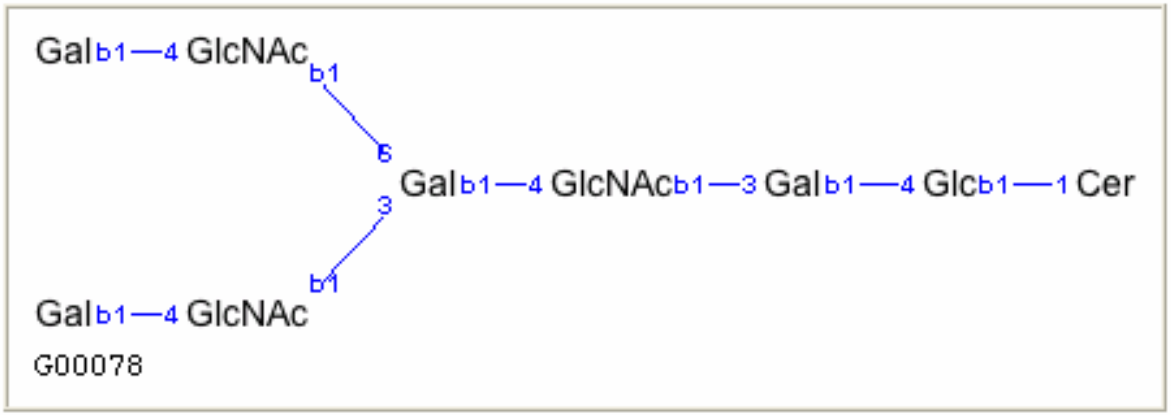
“View Structure”

KEGG Glycan Search

Compute Return

Query G00078

Structure



Database
- (•) KEGG GLYCAN
- ( ) CarbBank

Program
- (•) Gapped (Approximate match)
- ( ) Ungapped (Exact match)

Option
- (•) Global search
- ( ) Local search

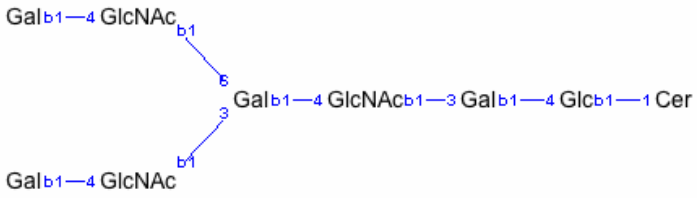
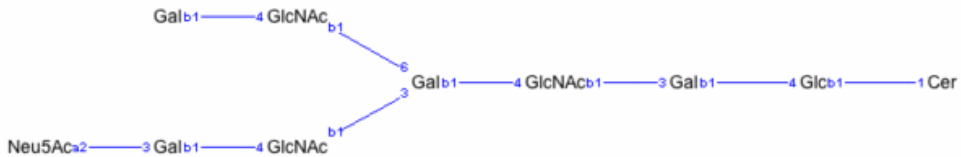
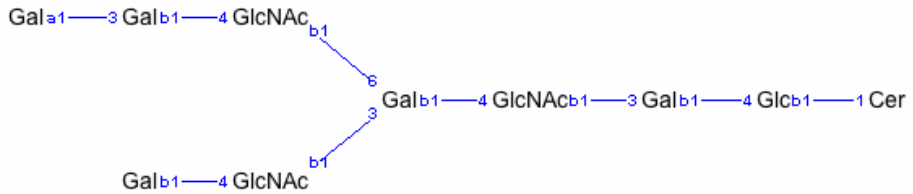
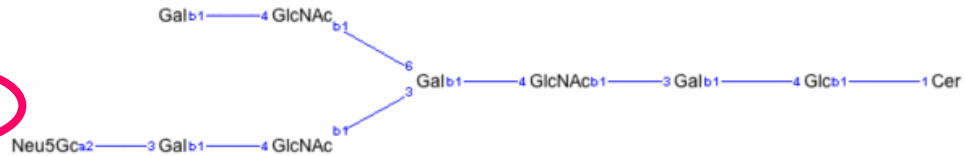
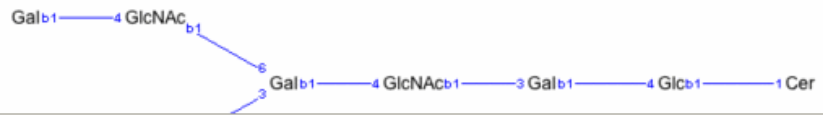
-> Show advanced options

Glycan Data Search Result

Top

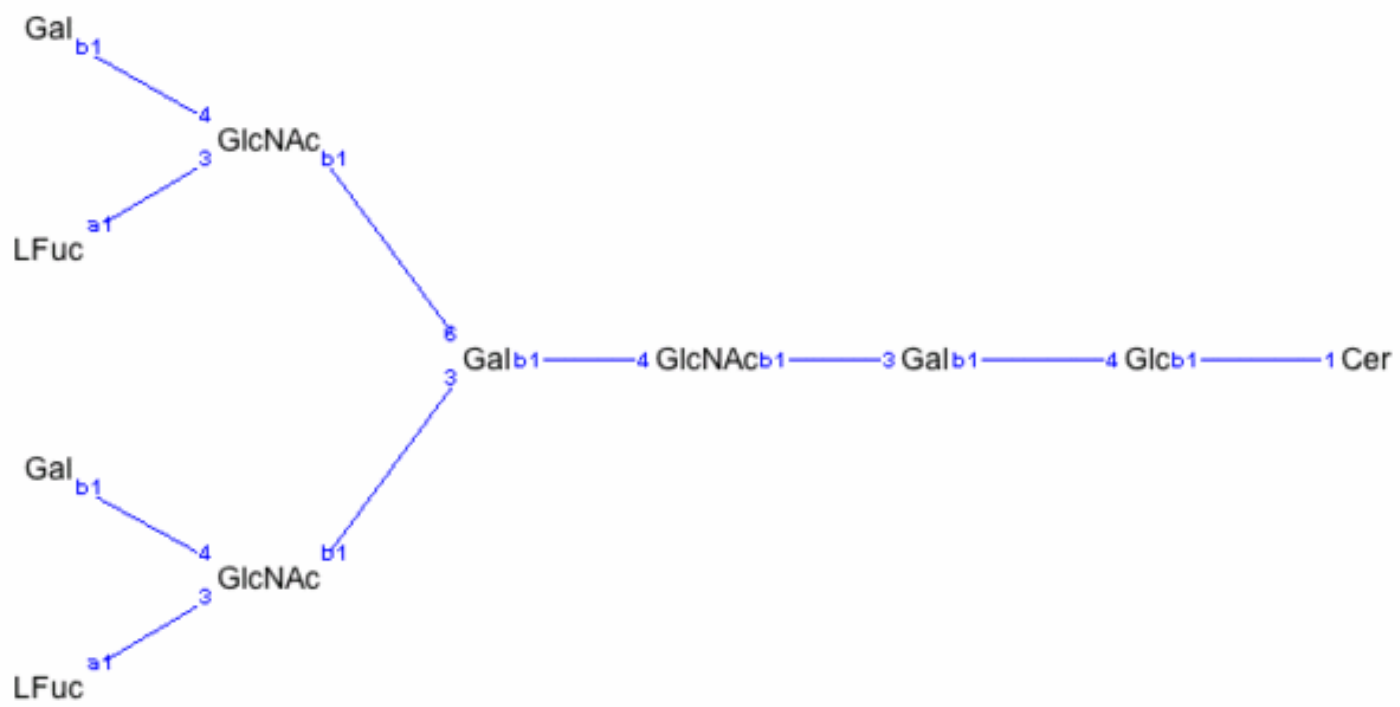
umber of entries in a page 20 Hide structure

ge : 1 Go of 36 Items : 1 - 20 of 720 Top Previous Next Bottom

No	Entry	Structure	Name	Composition	Class
	G00078	 G00078 Similarity-Score : 900	iso-nLc8Cer LacNAc-Lc6Cer I-antigen Lactoisooctaosylceramide	(Gal)4 (Glc)1 (GlcNAc)3 (Cer)1	Glycolipid; Sphingolipid
	G00879	 G00879 Similarity-Score : 800		(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Ac)1 (Cer)1	Glycolipid; Sphingolipid
	G02772	 G02772 Similarity-Score : 800		(Gal)5 (Glc)1 (GlcNAc)3 (Cer)1	Glycolipid; Sphingolipid
	G04167	 G04167 Similarity-Score : 800		(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Gc)1 (Cer)1	Glycolipid; Sphingolipid
	G04168			(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Ac)1	Glycolipid;

GLYCAN: G04450

Help

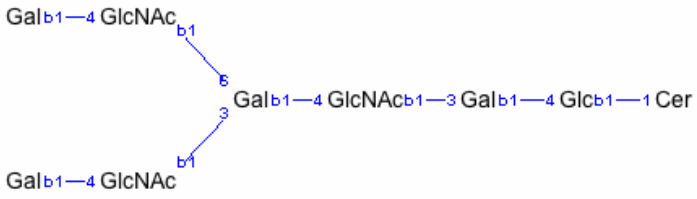
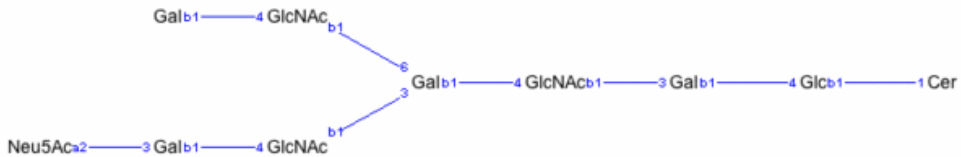
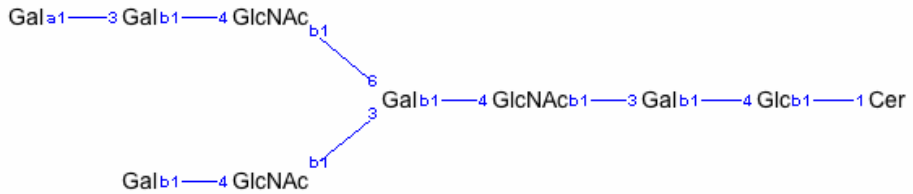
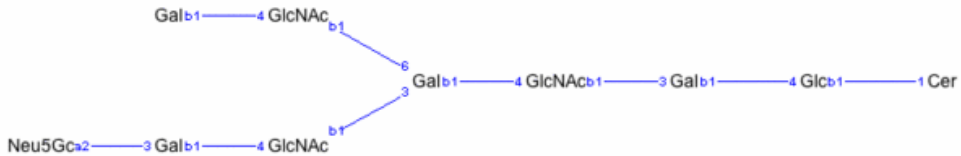
Entry	G04450 Glycan
Composition	(Gal)4 (Glc)1 (GlcNAc)3 (LFuc)2 (Cer)1
Mass	1712.6 (Cer)
Structure	 G04450 KCF file KCaM
Class	Glycolipid; Sphingolipid
Other DBs	CCSD: 20753
LinkDB	All DBs
KCF data	Show

Glycan Data Search Result

Top

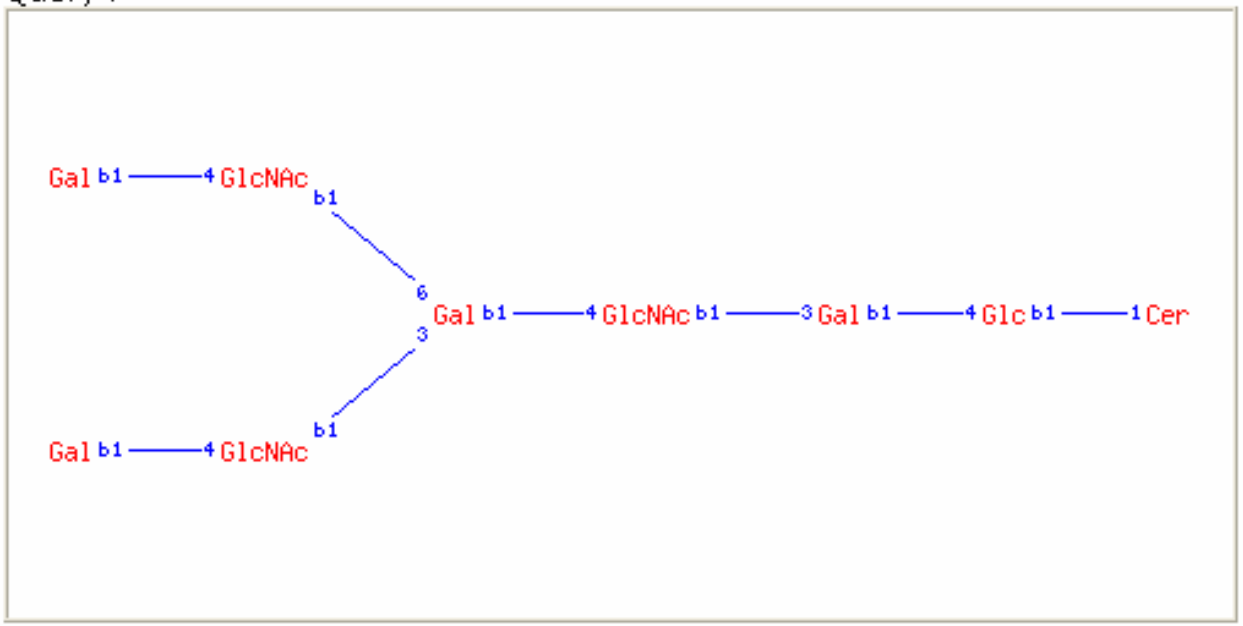
umber of entries in a page 20 Hide structure

ge : 1 Go of 36 Items : 1 - 20 of 720 Top Previous Next Bottom

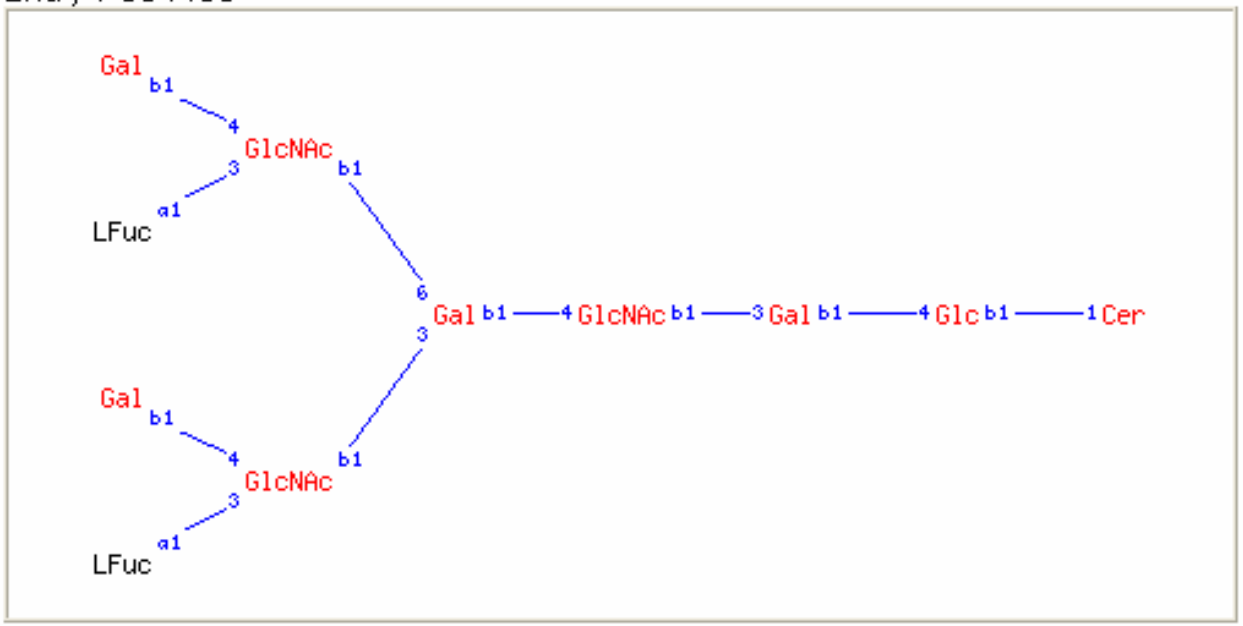
No	Entry	Structure	Name	Composition	Class
	G00078	 G00078 Similarity-Score : 900	iso-nLc8Cer LacNAc-Lc6Cer I-antigen Lactoisooctaosylceramide	(Gal)4 (Glc)1 (GlcNAc)3 (Cer)1	Glycolipid; Sphingolipid
	G00879	 G00879 Similarity-Score : 800		(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Ac)1 (Cer)1	Glycolipid; Sphingolipid
	G02772	 G02772 Similarity-Score : 800		(Gal)5 (Glc)1 (GlcNAc)3 (Cer)1	Glycolipid; Sphingolipid
	G04167	 G04167 Similarity-Score : 800		(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Gc)1 (Cer)1	Glycolipid; Sphingolipid
	G04168			(Gal)4 (Glc)1 (GlcNAc)3 (Neu5Ac)1	Glycolipid;

Similarity-Score : 700

Query :



Entry : G04450



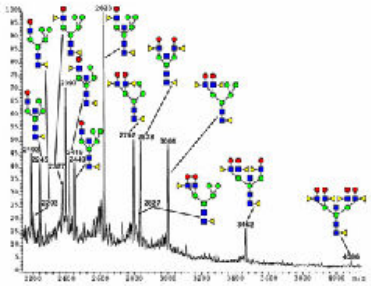
Consortium for Functional Glycomics (CFG)

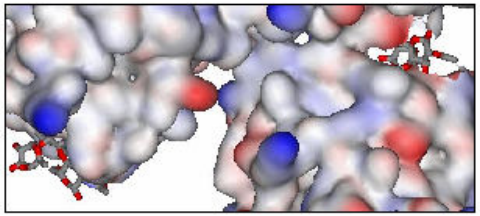
- Consortium home page: http://www.functionalglycomics.org/
- Consortium of major universities and research institutes worldwide
- Aim: to provide a central resource for glycomics research
- Also provides requested resources to promote participating investigators' research
  - Glycan arrays and data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

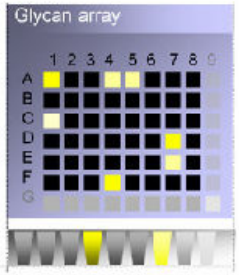
CONSORTIUM FOR FUNCTIONAL GLYCOMICS

funded by National Institute of General Medical Sciences

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

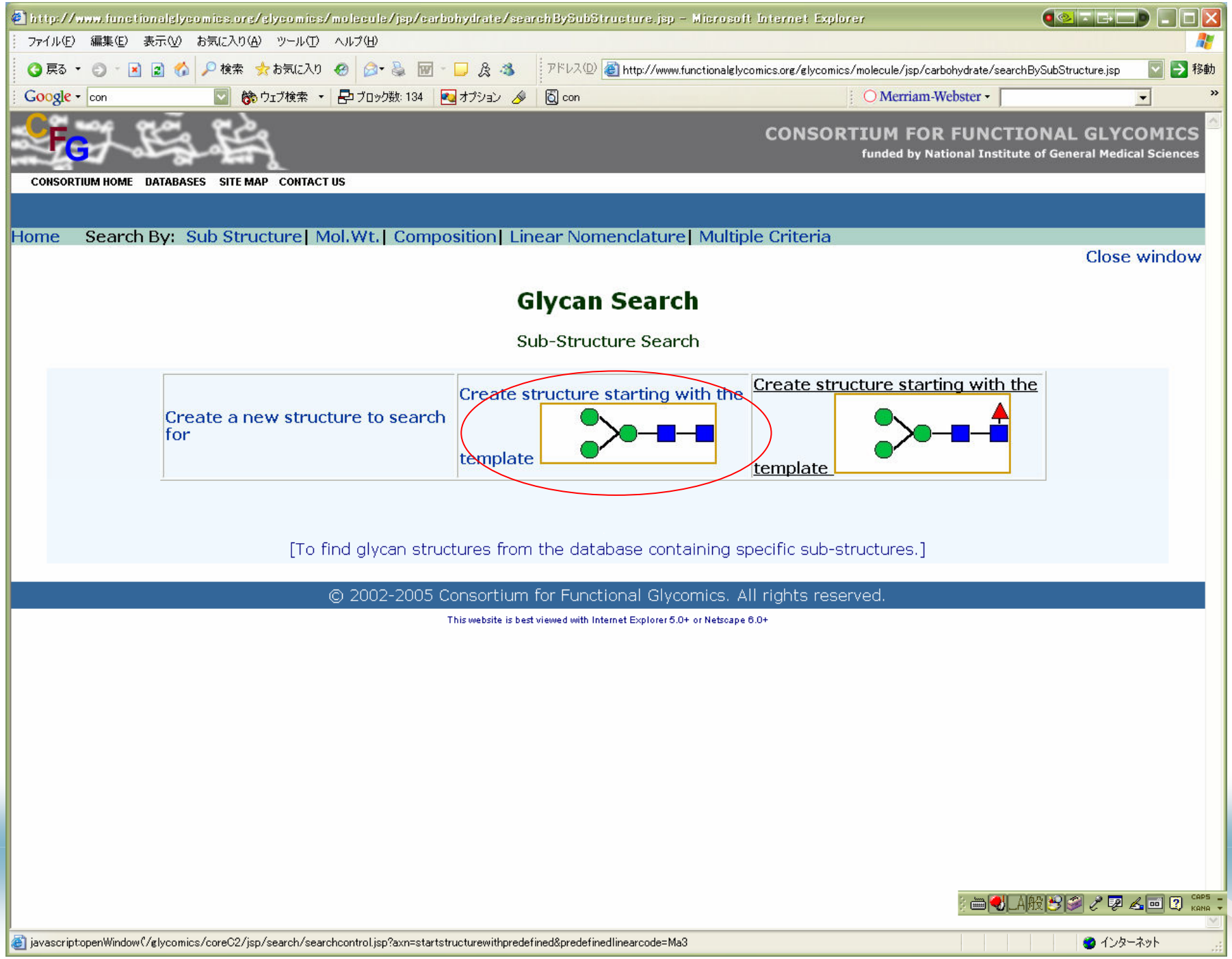
Source of glycan structures

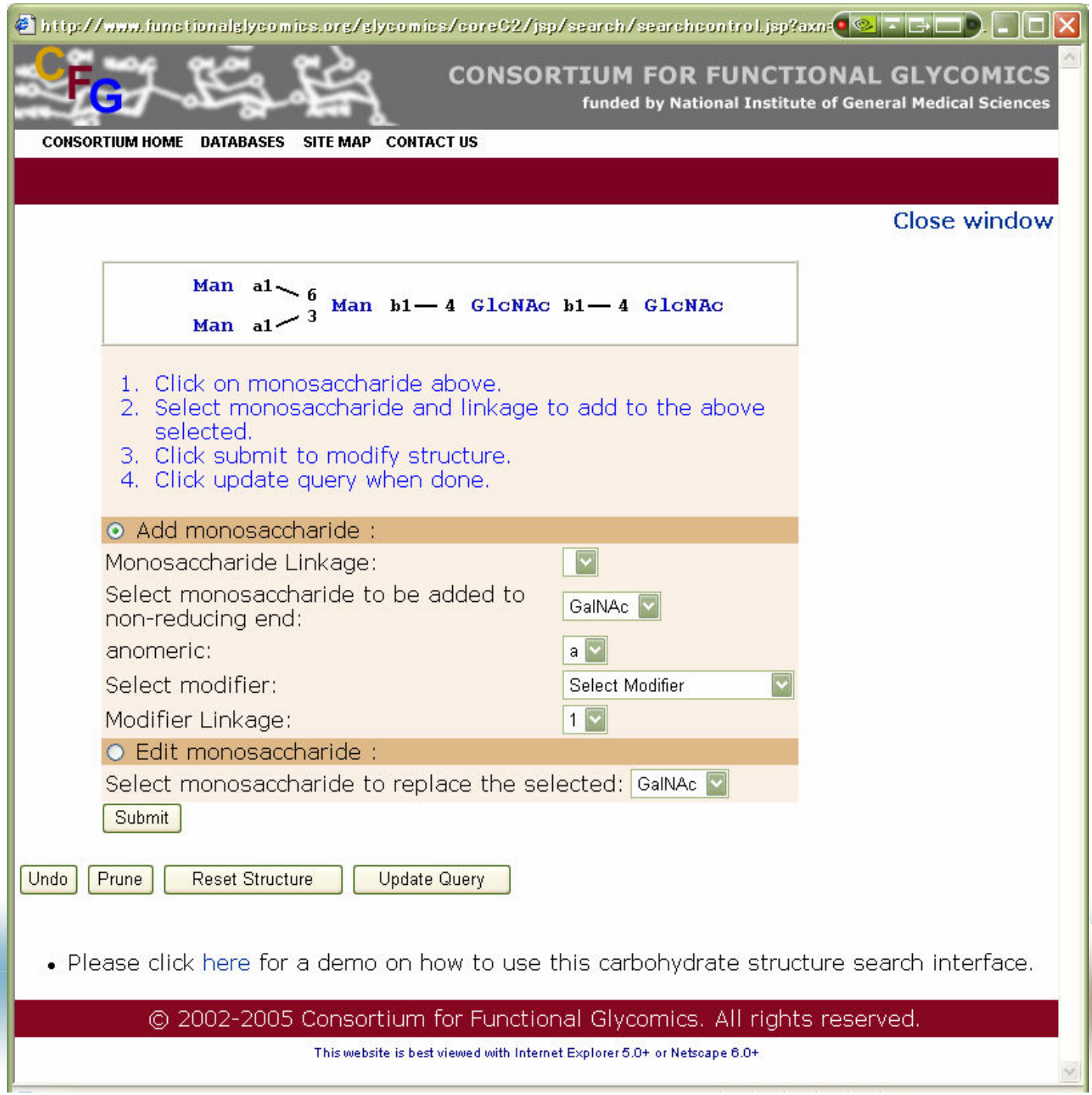
- *N*- and *O*-linked glycans from CarbBank
- Glycominds Ltd. seed database
- *N*- and *O*-linked glycans identified in tissues and cells analyzed by the Analytical Glycotechnology Core (C)
- Glycans elaborated on the glycan array
- Glycans synthesized by the Carbohydrate Synthesis Core (D) and available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

This website is best viewed with Internet Explorer 5.0+ or Netscape 6.0+


http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/searchBySubStructure.jsp - Microsoft Internet Explorer
CONSORTIUM FOR FUNCTIONAL GLYCOMICS
funded by National Institute of General Medical Sciences
CONSORTIUM HOME DATABASES SITE MAP CONTACT US
Home Search By: Sub Structure| Mol.Wt.| Composition| Linear Nomenclature| Multiple Criteria
Close window
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 5.0+ or Netscape 6.0+


http://www.functionalglycomics.org/glycomics/coreC2/jsp/search/searchcontrol.jsp?axn
CONSORTIUM FOR FUNCTIONAL GLYCOMICS
funded by National Institute of General Medical Sciences
CONSORTIUM HOME
DATABASES
SITE MAP
CONTACT US
Close window
Man a1 6
Man a1 3
Man b1 4 GlcNAc b1 4 GlcNAc
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:
Select monosaccharide to be added to non-reducing end:
GalNAc
anomeric:
a
Select modifier:
Select Modifier
Modifier Linkage:
1
Edit monosaccharide :
Select monosaccharide to replace the selected:
GalNAc
Submit
Undo
Prune
Reset Structure
Update Query
Please click here for a demo on how to use this carbohydrate structure search interface.
© 2002-2005 Consortium for Functional Glycomics. All rights reserved.
This website is best viewed with Internet Explorer 5.0+ or Netscape 6.0+

CONSORTIUM FOR FUNCTIONAL GLYCOMICS

funded by National Institute of General Medical Sciences

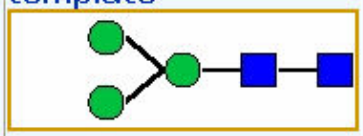
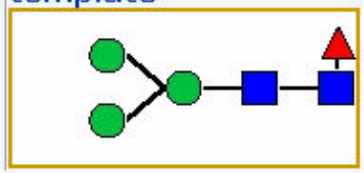
CONSORTIUM HOME DATABASES SITE MAP CONTACT US

HomeSearch By: Sub Structure| Mol.Wt.| Composition| Linear Nomenclature| Multiple Criteria

Close window

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 a1—6
Man a1—3 Man b1—4 GlcNAc b1—4 GlcNAc

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

Submit

This website is best viewed with Internet Explorer 5.0+ or Netscape 6.0+

Glycan Structures Database

Interactive Search Results [Help]

Original parameters used to obtain the full results below:

Further refinements:

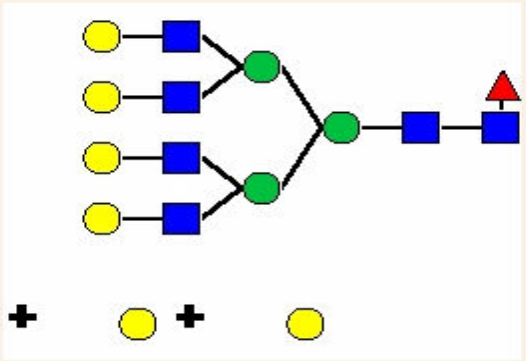
- Original Results

Results 1 to 26 of 2509 results

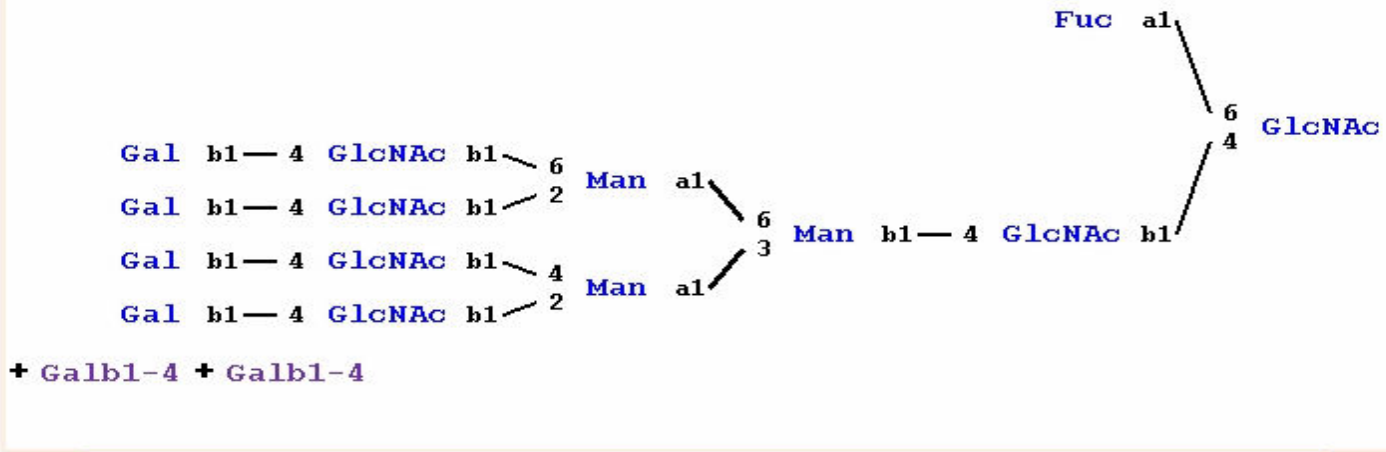
Oligosaccharide Molecular Wt.	IUPAC	Composition	Family	Sub Family	Source
2842.59	Galb1-4=2%|Galb1-4=1%|2%1%Galb1-4GlcNAcb1-2(2%1%Galb1-4GlcNAcb1-4)Mana1-3(2%1%Galb1-4GlcNAcb1-2(2%1%Galb1-4GlcNAcb1-6)Mana1-6)Manb1-4GlcNAcb1-4(Fuca1- 6)GlcNAc	$HEX_9DHEX_1HEXNAC_6$	N-linked	Complex	• Sus scrofa(Pig):Ovary:Follicle cells
2735.49	NeuAca2-3Galb1-4GlcNAcb1-2(Galb1-4GlcNAcb1-4)Mana1 -3(NeuAca2-3Galb1-4GlcNAcb1-2Mana1-6)Manb1-4GlcNAc b1-4(Fuca1-6)GlcNAc	$HEX_6DHEX_1HEXNAC_5NEUAC_2$	N-linked	Complex	• Cricetulus griseus(Chineese Hamster):CHO cells
5360.97	GlcNAcb1-3Galb1-4GlcNAcb1-3=1%|1%GlcNAcb1-3Galb1-4GlcNAcb1-3Galb1-4GlcNAcb1-2(1%GlcNAcb1-3Galb1-4GlcNAcb1-3Galb1-4GlcNAcb1-4)Mana1-3(1%GlcNAcb1-3Galb1 -4GlcNAcb1-3Galb1-4GlcNAcb1-2(1%GlcNAcb1-3Galb1-4GlcNAcb1-3Galb1-4GlcNAcb1-6)Mana1-6)Manb1-4GlcNAcb1 -4(Fuca1-6)GlcNAc	$HEX_{12}DHEX_1HEXNAC_{16}$	N-linked	Complex	• Sus scrofa(Pig):Ovary:Follicle cells
1422.29	Mana1-3(Galb1-4GlcNAcb1-2Mana1-6)Manb1-4GlcNAcb1-4(Fuca1-6)GlcNAc;[NP]	$HEX_4DHEX_1HEXNAC_3$	N-linked	Complex	
	NeuAca2-3Galb1-4GlcNAcb1-2(NeuAca2-3Galb1-4GlcNAcb 1-6)Mana1-3(NeuAca2-				

Glycan : carbNlink_21394_P

Cartoon Representation



IUPAC 2D Representation



IUPAC Code

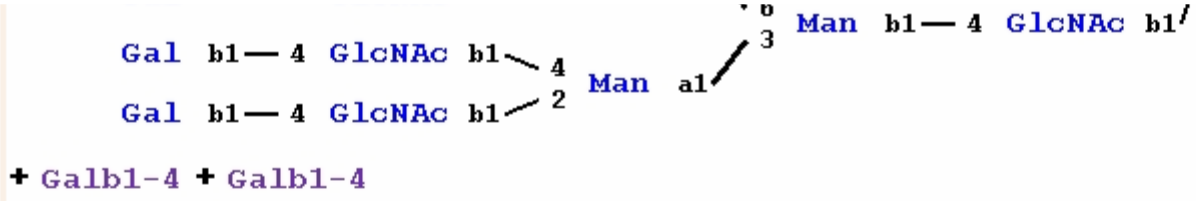
Gal b1-4=2%| *Gal* b1-4=1%|2% 1% *Gal* b1-4 *GlcNAc* b1-2(2% 1% *Gal* b1-4 *GlcNAc* b1-4) *Man* a1-3(2% 1% *Gal* b1-4 *GlcNAc* b1-2(2% 1% *Gal* b1-4 *GlcNAc* b1-6) *Man* a1-6) *Man* b1-4 *GlcNAc* b1-4(*Fuc* a1-6) *GlcNAc*

Linear Code

Ab4=2%|Ab4=1%|2%1%Ab4GNb2(2%1%Ab4GNb4)Ma3(2%1%Ab4GNb2(2%1%Ab4GNb6)Ma6) Mb4GN b4(Fa6)GN

General Information

Glycan Family:	N-linked
Sub. Family:	Complex



IUPAC Code

Gal b1-4=2%| *Gal* b1-4=1%|2% 1% *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* a1-6) *Man* b1-4 *GlcNAc* b1-4(*Fuc* a1-6) *GlcNAc*

Linear Code

Ab4=2%|Ab4=1%|2%1%Ab4GNb2(2%1%Ab4GNb4)Ma3(2%1%Ab4GNb2(2%1%Ab4GNb6)Ma6) Mb4GN b4(Fa6)GN

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:	$dHex_1$ $HexNAc_6$ Hex_9
Status:	Public

References

- Mori E, Takasaki S, Hedrick JL, Wardrip NJ, Mori T, Kobata A Biochemistry 1991;[2078-2087] {*PubMed*}

Biological Sources

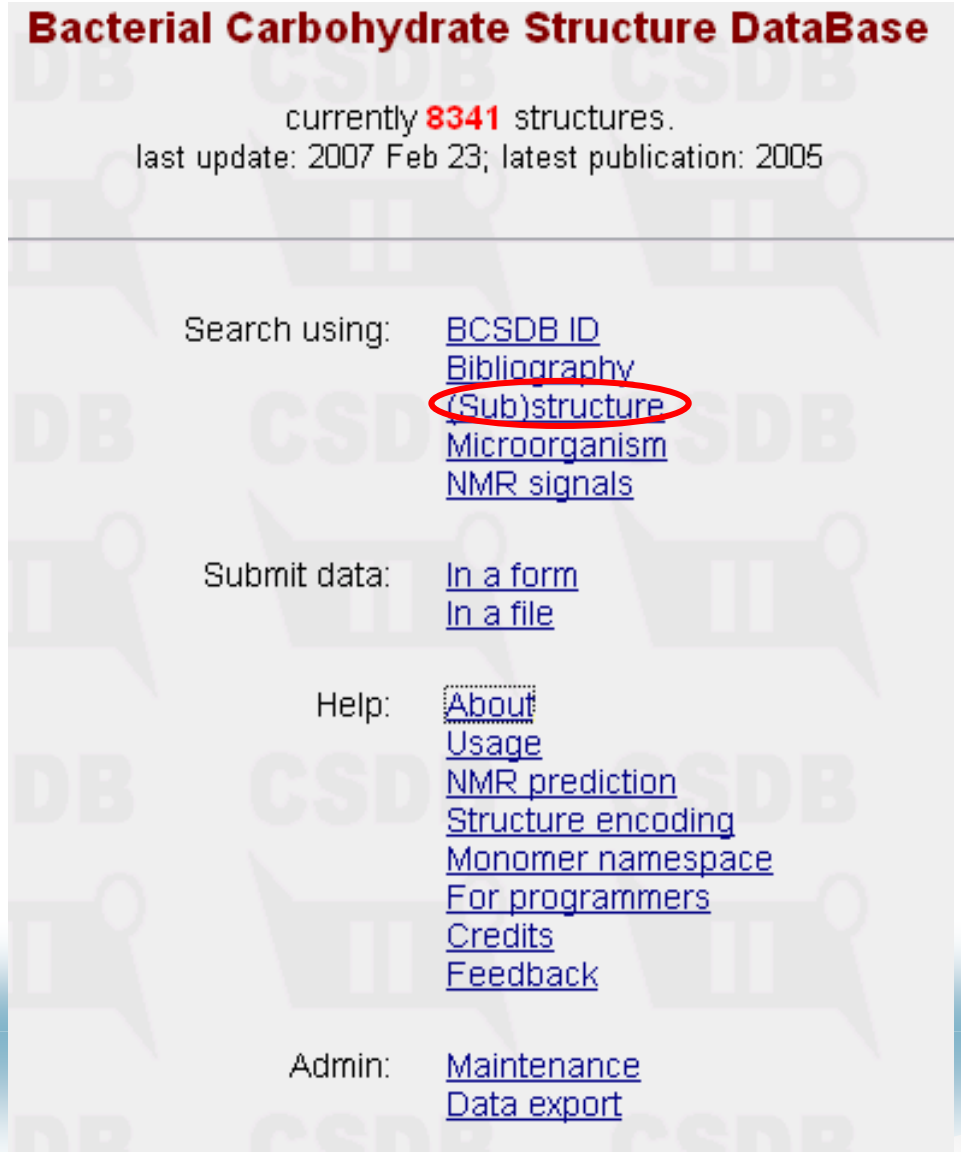
Taxonomy Name	Organ	Tissue Type	Cell Type
Sus scrofa(Pig)	Ovary		Follicle cells

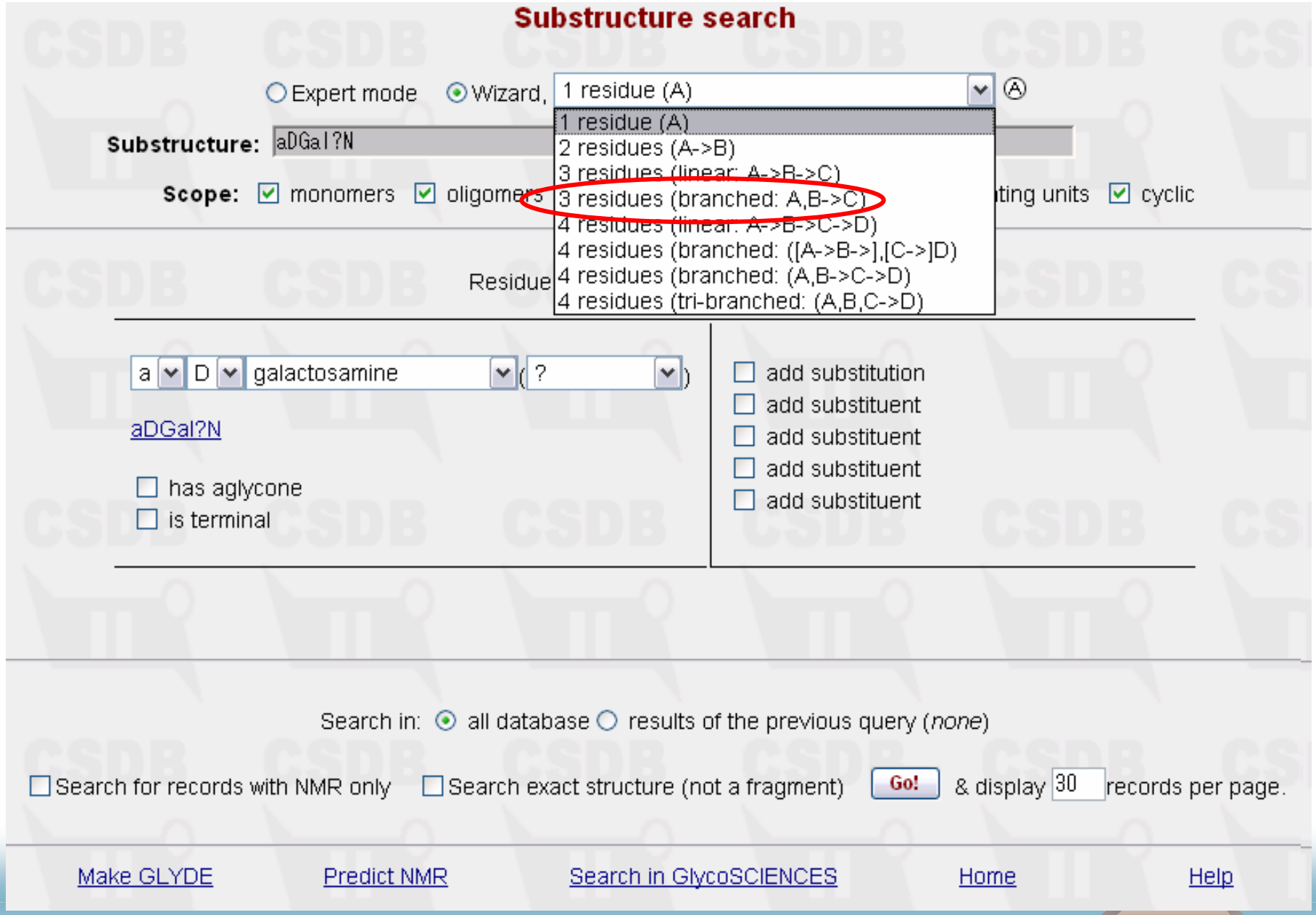
BCSDB: Bacterial Carbohydrate Structure DataBase

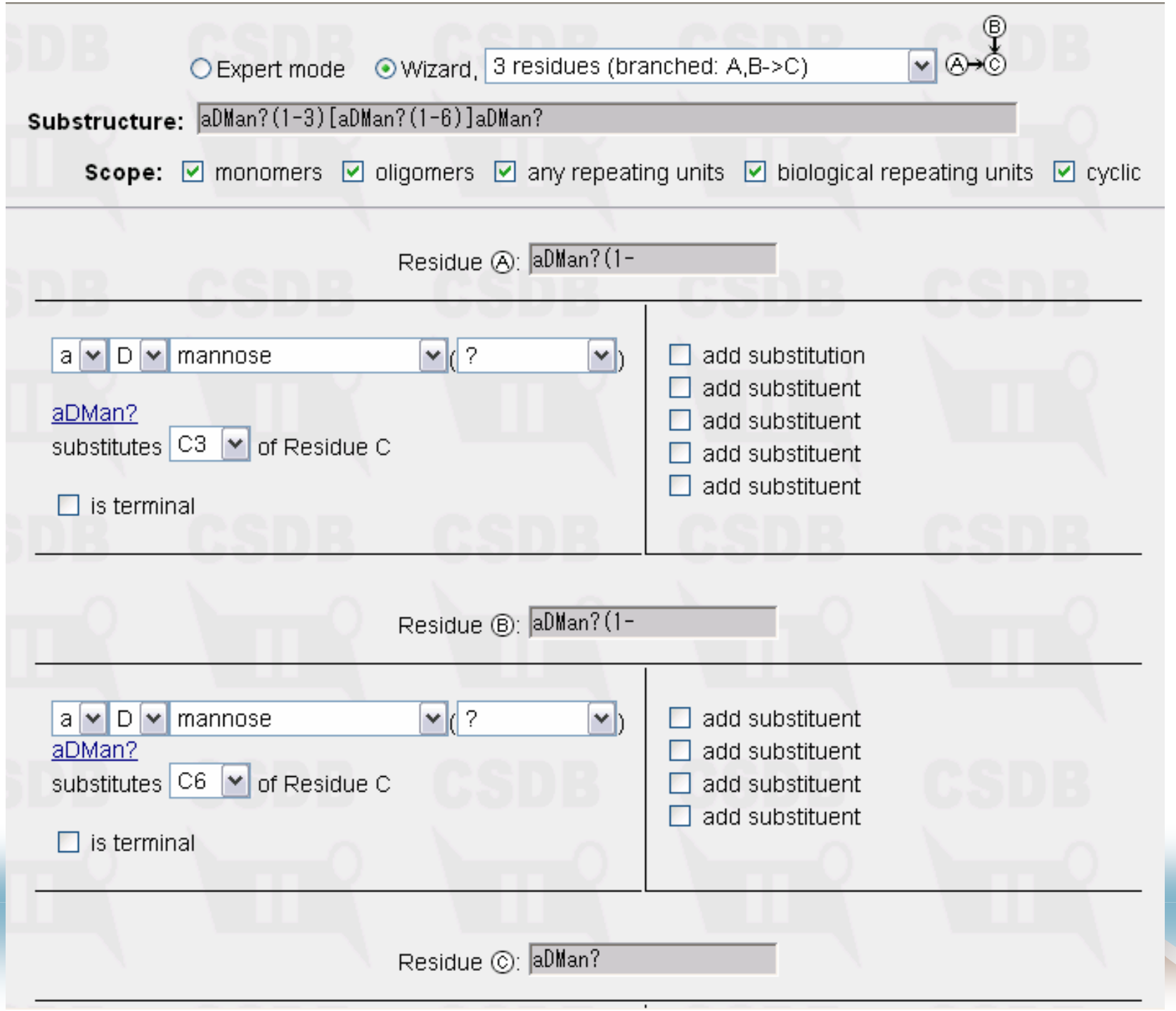
http://www.glyco.ac.ru/bcsdb/start.shtml

- 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


Substructure search
Expert mode
Wizard,
1 residue (A)
1 residue (A)
2 residues (A->B)
3 residues (linear: A->B->C)
3 residues (branched: A,B->C)
4 residues (linear: A->B->C->D)
4 residues (branched: ([A->B->],[C->]D)
4 residues (branched: (A,B->C->D)
4 residues (tri-branched: (A,B,C->D)
Substructure:
aDGal?N
Scope:
monomers
oligomers
cyclic
Residue
a
D
galactosamine
?
aDGal?N
has aglycone
is terminal
add substitution
add substituent
add substituent
add substituent
add substituent
Search in:
all database
results of the previous query (none)
Search for records with NMR only
Search exact structure (not a fragment)
Go!
& display
30
records per page.
Make GLYDE
Predict NMR
Search in GlycoSCIENCES
Home
Help

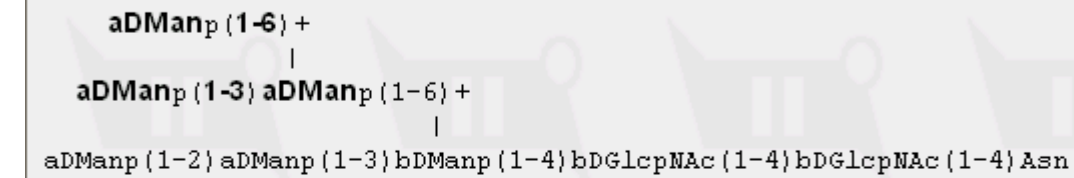

Expert mode
Wizard,
3 residues (branched: A,B->C)
Substructure:
aDMan?(1-3)[aDMan?(1-6)]aDMan?
Scope:
monomers
oligomers
any repeating units
biological repeating units
cyclic
Residue Ⓐ:
aDMan?(1-
a
D
mannose
?
add substitution
add substituent
add substituent
add substituent
add substituent
aDMan?
substitutes
C3
of Residue C
is terminal
Residue Ⓑ:
aDMan?(1-
a
D
mannose
?
add substituent
add substituent
add substituent
add substituent
aDMan?
substitutes
C6
of Residue C
is terminal
Residue Ⓒ:
aDMan?

Found **56** records of **8341**
Displayed records from **1** to **30**
Next 30 records
Expand all records

---

1. (*BCSDB ID:* 111464)

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



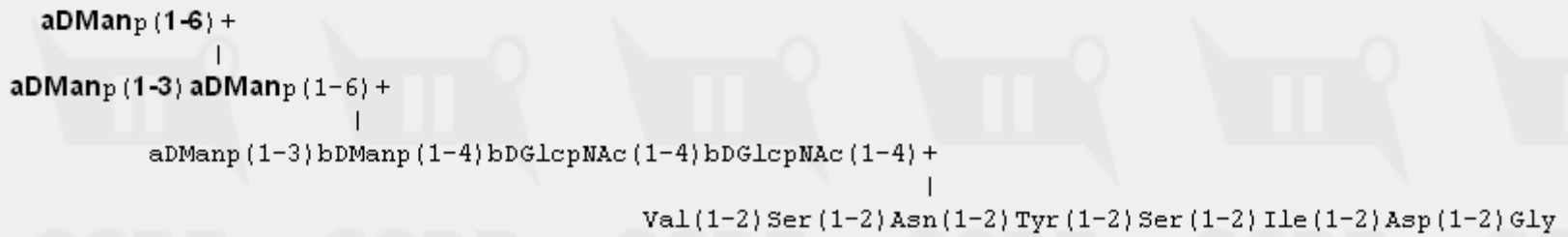
***Aspergillus niger* IFO31125** (NCBI Taxonomy)

Expand this record

---

2. (*BCSDB ID:* 129098)

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



Aspergillus oryzae (NCBI Taxonomy)

Found **56** records of **8341**
Displayed records from **1** to **30**
Next 30 records
Expand all records

1. (*BCSDB ID:* 111464)

Akiba S, Yamamoto K, Kumagai H
Effects of size of carbohydrate chain on protease digestion o
Bioscience, Biotechnology, and Biochemistry **59** (1995) 1048-1051

aDManp(**1-6**)+
|
aDManp(**1-3**)**aDMan**p(1-6)+
|
aDManp(1-2)aDManp(1-3)bDManp(1-4)bDGlcpNAc(1-4)bDGl

***Aspergillus niger* IFO31125** (NCBI Taxonomy)

Taxonomic group: fungi *(Phylum: Ascomycota)*

Structure type: oligomer
Compound class: N-linked glycoprotein

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

NCBI Taxonomy refs (TaxIDs): 5061

Make GLYDE 1.2 description

Predict ^{13}C NMR assignment table

Find this structure in GlycoSCIENCES DB

Collapse this record

2. (*BCSDB ID:* 129098)

GLYDE 1.2 representation of the structure is:

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE glycanset PUBLIC "Glyde 1.2" "http://www.glyco.ac.ru/bcsdb/help/glyde12.dtd">
<glycan_set xmlns="">
  <glycan>
    <residue name="Asn" anomer="null" chirality="null" ring_form="null" linking_atom="null" link="null" type="reducing">
      <residue name="GlcNAc" anomer="b" chirality="D" ring_form="p" linking_atom="1" link="4" type="glycosyl">
        <residue name="GlcNAc" anomer="b" chirality="D" ring_form="p" linking_atom="1" link="4" type="glycosyl">
          <residue name="Man" anomer="b" chirality="D" ring_form="p" linking_atom="1" link="4" type="glycosyl">
            <residue name="Man" anomer="a" chirality="D" ring_form="p" linking_atom="1" link="3" type="glycosyl">
              <residue name="Man" anomer="a" chirality="D" ring_form="p" linking_atom="1" link="2" type="glycosyl"/>
            </residue>
            <residue name="Man" anomer="a" chirality="D" ring_form="p" linking_atom="1" link="6" type="glycosyl">
              <residue name="Man" anomer="a" chirality="D" ring_form="p" linking_atom="1" link="3" type="glycosyl"/>
              <residue name="Man" anomer="a" chirality="D" ring_form="p" linking_atom="1" link="6" type="glycosyl"/>
            </residue>
          </residue>
        </residue>
      </residue>
    </residue>
  </glycan>
</glycan_set>
```

BCSDB representation of the structure is:

aDManp(1-2)aDManp(1-3)[aDManp(1-3)[aDManp(1-6)]aDManp(1-6)]bDManp(1-4)[Ac(1-2)]bDGlcpN(1-4)[Ac(1-2)]bDGlcpN(1-4)xXAsn

Close window

aDManp(1-2)aDManp(1-3)[aDManp(1-3)[aDManp(1-6)]aDManp(1-6)]bDManp(1-4)[Ac(1-2)]bDGlcpN(1-4)[Ac(1-2)]bDGlcpN(1-4)xXAsn

Simulated ^{13}C NMR spectrum (in D_2O) is:

^{13}C NMR data:

Linkage	Residue	C1	C2	C3	C4	C5	C6	Accuracy
0	Asn	52.7	174.2	35.1	180.5			*1*
4	b-D-GlcNAc	102.2	57.0	73.7	80.2	75.9	61.3	*4*
4,4	b-D-GlcNAc	102.5	57.0	73.7	80.2	75.9	61.3	*4*
4,4,4	b-D-Man	101.2	71.3	81.7	67.3	75.6	66.8	*2*
4,4,4,3	a-D-Man	103.6	79.3	70.8	67.8	73.4	61.9	*4*
4,4,4,3,2	a-D-Man	103	71.5	71.5	68.2	74.2	62.3	*4*
4,4,4,6	a-D-Man	100.8	70.8	78.9	67.6	72.9	66.8	*2*
4,4,4,6,3	a-D-Man	103.6	71.5	71.5	68.2	74.2	62.3	*4*
4,4,4,6,6	a-D-Man	100.8	71.5	71.5	68.2	74.2	62.3	*4*

Click here to show free residues spectra and effects applied

Overall prediction consistency: 3.22 (0 = worst; 4 = best. Details here)

Sorted ^{13}C NMR spectrum:
35.1 52.7 57.0 57.0 61.3 61.3 61.9 62.3 62.3 62.3 66.8 66.8 67.3 67.6 67.8 68.2 68.2 68.2 70.8 70.8 71.3 71.5 71.5 71.5 71.5 71.5 71.5 72.9 73.4 73.7 73.7 74.2 74.2 74.2 75.6 75.9 75.9 78.9 79.3 80.2 80.2 81.7 100.8 100.8 101.2 102.2 102.5 103 103.6 103.6 174.2 180.5

The signals of acetyl groups and other monovalent substituents are omitted. This spectrum also does not contain signals for the residues with error message in the assignment table instead of chemical shifts.

You may use the following data to refer to the ^{13}C NMR spectra simulation engine used to obtain these results:
F.V. Toukach "Computer-assisted structural analysis of glycopolymers" (Eurocarb-12 proceedings, 2003, p. PA-004)

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.

```
      a-D-Manp-(1-6)+
                    |
              a-D-Manp-(1-6)+
                    |       |
      a-D-Manp-(1-3)+   b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc-(1-4)-Asn
                                |
a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

Molecular weight: 1495
Chemical formula: $C_{56}H_{94}N_4O_{42}$

^{1}H NMR data:
Recorded in $_{300}$ at 0.0 K
Spectrometer: 400.0 MHz

Residue	Linkage	Atom	δ, ppm	J, from	J, to	J, Hz	
b-D-GlcpNAc	4	H1	5.04			0.0	
b-D-GlcpNAc	4	NAC	2.01			0.0	a
b-D-GlcpNAc	4	NAC	2.01			0.0	a
b-D-GlcpNAc	4,4	H1	4.60			0.0	
b-D-GlcpNAc	4,4	NAC	2.06			0.0	
b-D-Manp	4,4,4	H1	4.77			0.0	
b-D-Manp	4,4,4	H2	4.23			0.0	
a-D-Manp	6,4,4,4	H1	4.87			0.0	
a-D-Manp	6,4,4,4	H2	4.14			0.0	
a-D-Manp	6,6,4,4,4	H1	4.91			0.0	
a-D-Manp	6,6,4,4,4	H2	3.98			0.0	
a-D-Manp	3,6,4,4,4	H1	5.10			0.0	
a-D-Manp	3,6,4,4,4	H2	4.07			0.0	

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

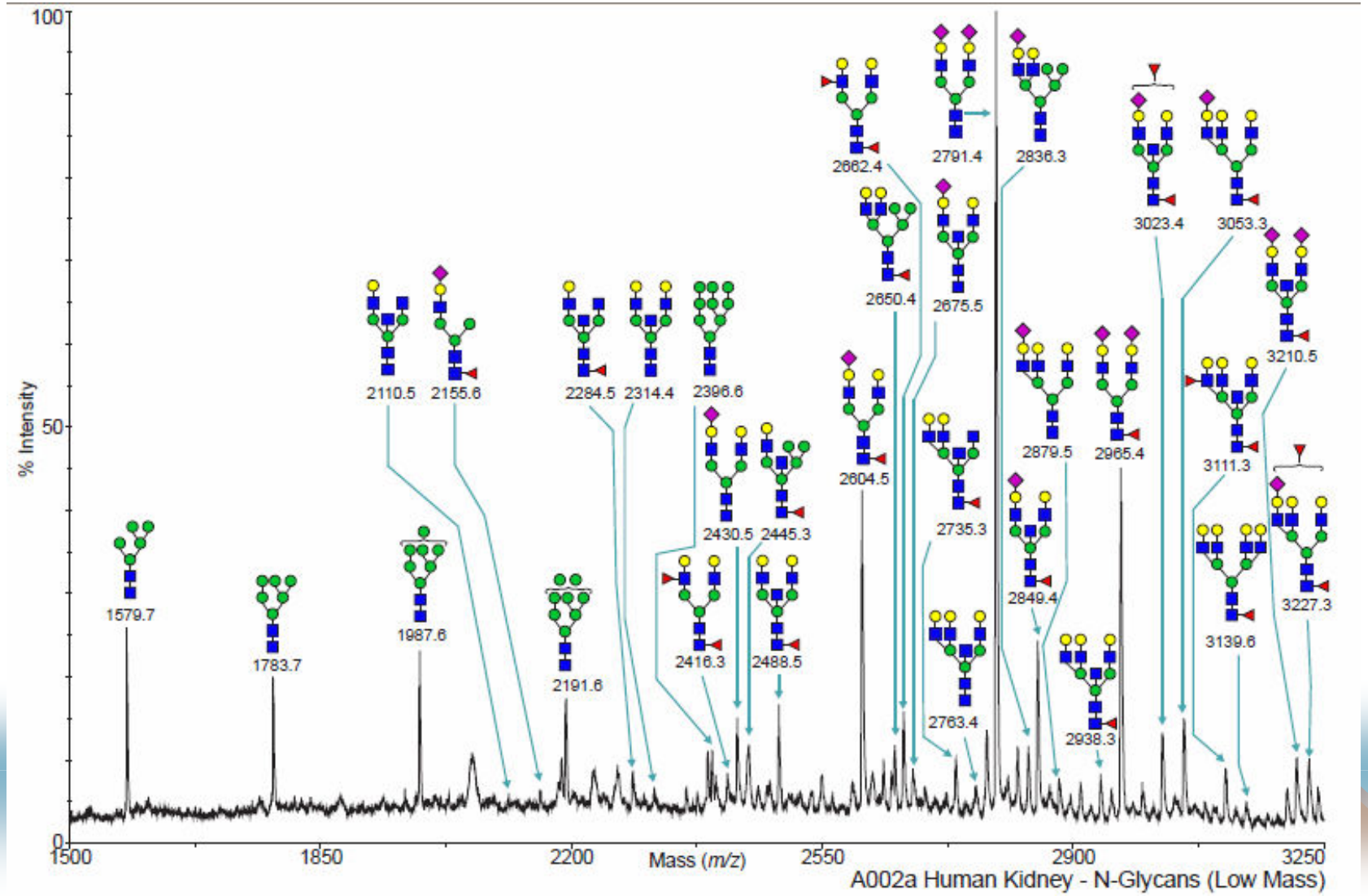
Current glycome informatics

- 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)
  - STAT (Gaucher et al, Anal. Chem., 2000)
  - 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)

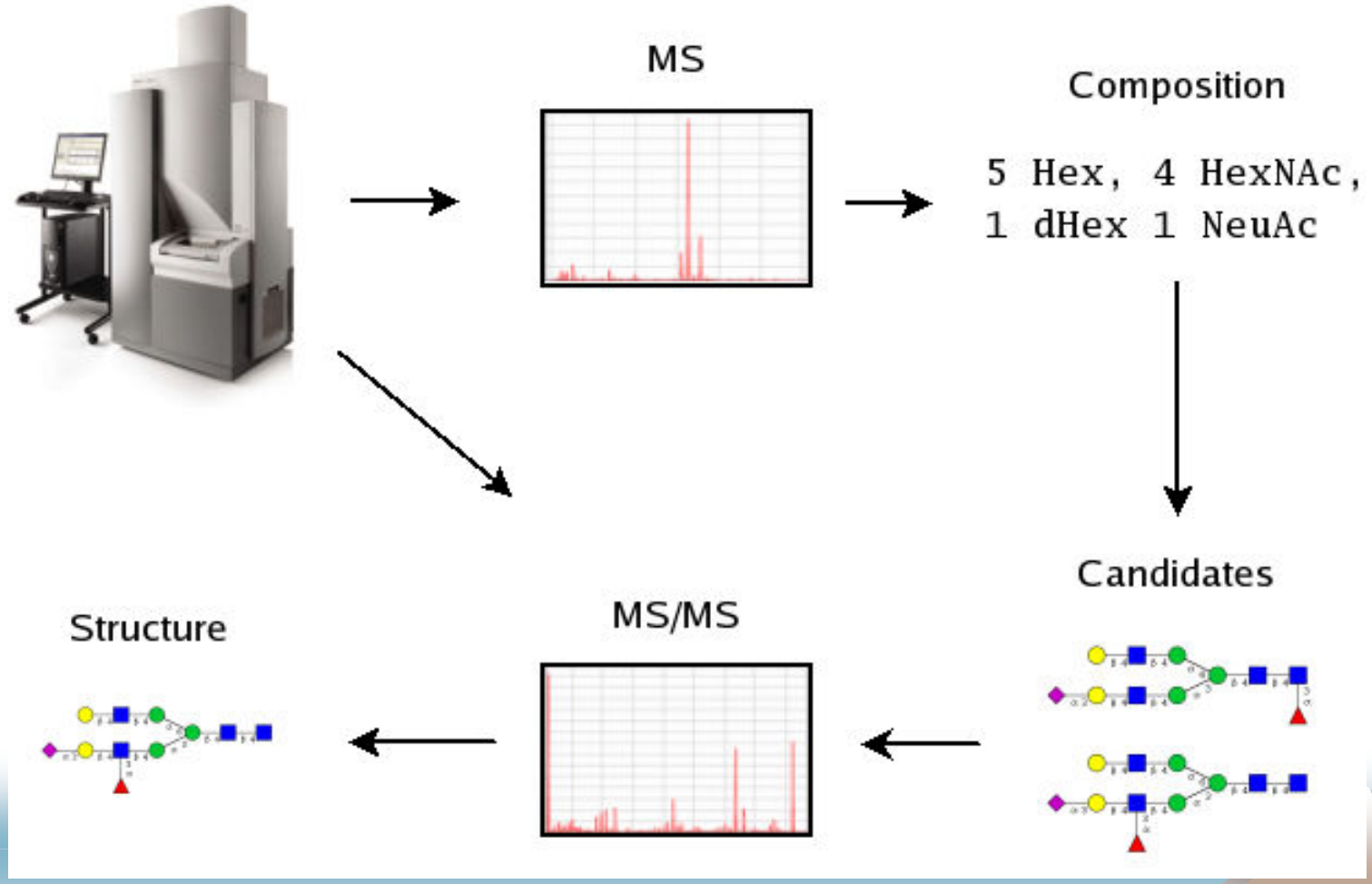
Automated Annotation of Mass Spectrometry Data



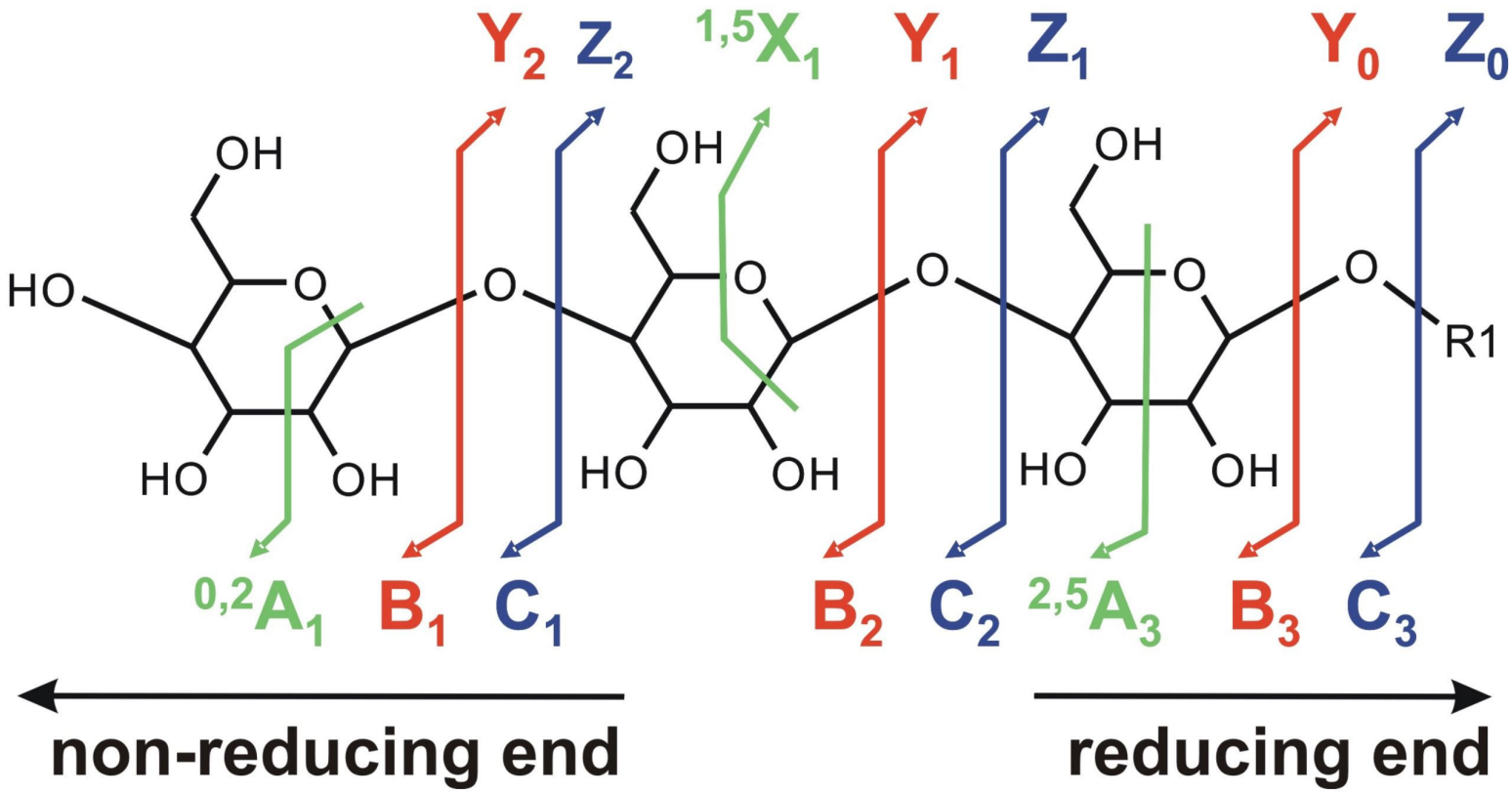
GlycoMod

- http://www.expasy.ch/tools/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

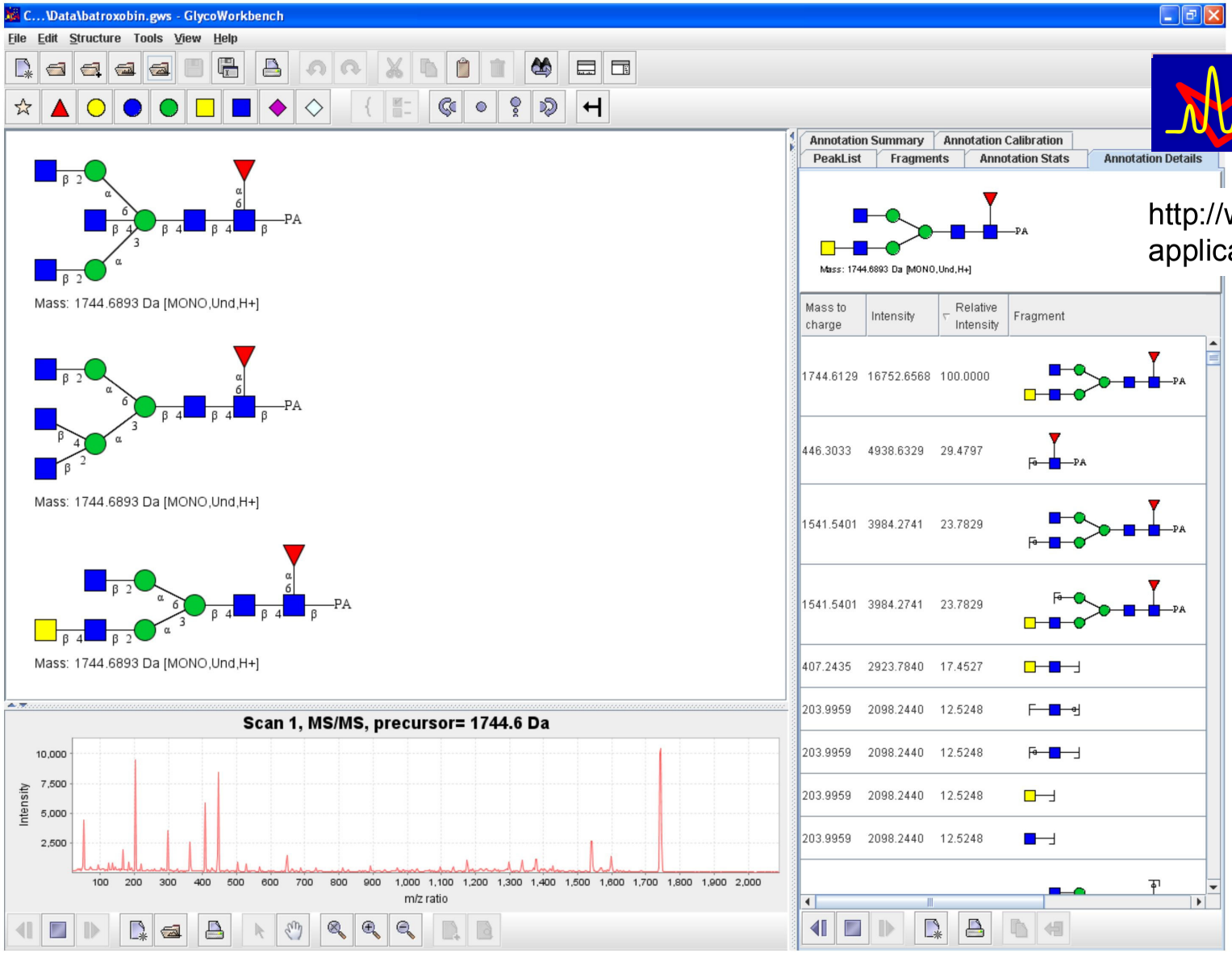
Experimental workflow for (semi-)automatic determination of glycan structures from raw data to fully assigned spectrum via composition analysis (GlycoPeakFinder) and fragment matching (GlycoWorkbench).



Nomenclature of MS fragments of carbohydrates as defined by Domon and Costello

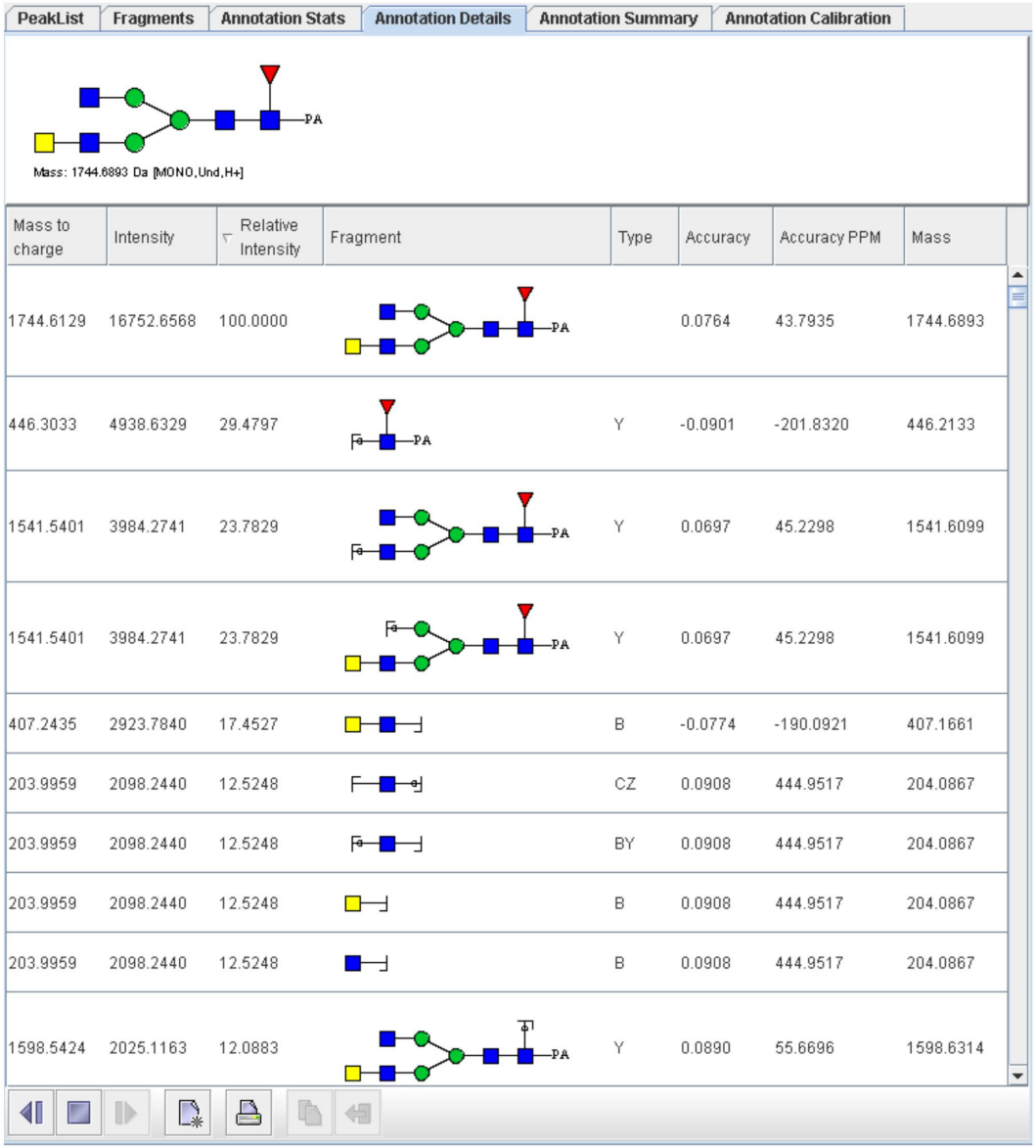


GlycoWorkbench
MS: Annotation of fragments



http://www.eurocarbdb.org/applications

GlycoWorkbench
MS: Annotation of fragments



Mass to charge	Intensity	Relative Intensity	Fragment	Type	Accuracy	Accuracy PPM	Mass
1744.6129	16752.6568	100.0000			0.0764	43.7935	1744.6893
446.3033	4938.6329	29.4797		Y	-0.0901	-201.8320	446.2133
1541.5401	3984.2741	23.7829		Y	0.0697	45.2298	1541.6099
1541.5401	3984.2741	23.7829		Y	0.0697	45.2298	1541.6099
407.2435	2923.7840	17.4527		B	-0.0774	-190.0921	407.1661
203.9959	2098.2440	12.5248		CZ	0.0908	444.9517	204.0867
203.9959	2098.2440	12.5248		BY	0.0908	444.9517	204.0867
203.9959	2098.2440	12.5248		B	0.0908	444.9517	204.0867
203.9959	2098.2440	12.5248		B	0.0908	444.9517	204.0867
1598.5424	2025.1163	12.0883		Y	0.0890	55.6696	1598.6314

EUROCarbDB

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 supertree 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)

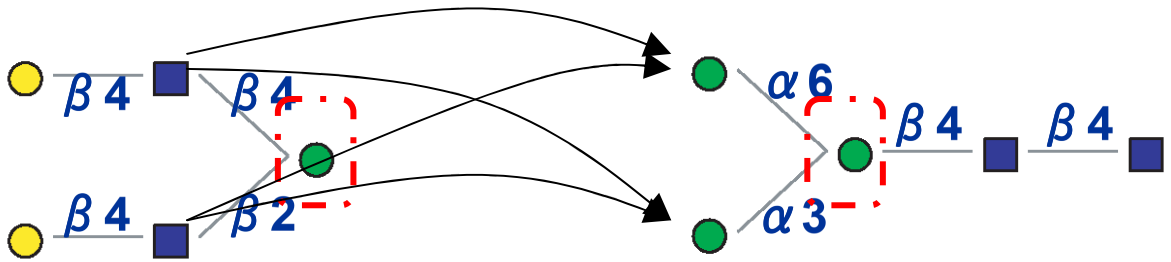
$$
\begin{aligned}
S[i,0] &= d \cdot i, \\
S[0,j] &= d \cdot j, \\
S[i,j] &= \max \begin{cases} S[i,j-1]+d, \\ S[i-1,j]+d, \\ S[i-1,j-1]+w(x_i,y_j) \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
S[i,0] &= 0, \\
S[0,j] &= 0, \\
S[i,j] &= \max \begin{cases} 0, \\ S[i,j-1]+d, \\ S[i-1,j]+d, \\ S[i-1,j-1]+w(x_i,y_j), \end{cases}
\end{aligned}
$$

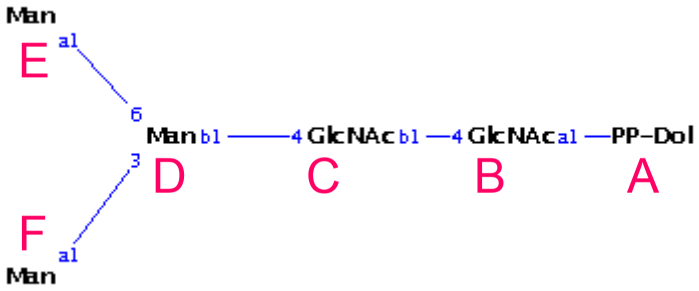
KCaM: KEGG Carbohydrate Matcher

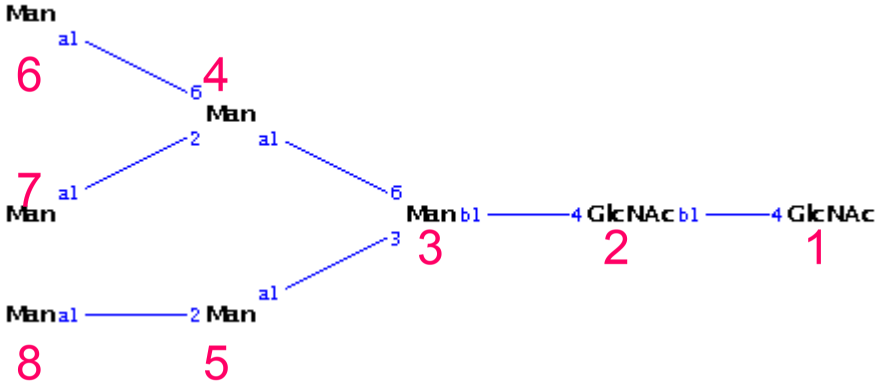
- Maximum Common Subtree Algorithm

$$
\begin{aligned}
R[u,0] &= 0, \\
R[0,v] &= 0, \\
R[u,v] &= 1 + \max_{\psi \in M(u,v)} \left\{ \sum_{u_i \in son\,s(u)} R[u_i, \psi(u_i)] \right\}
\end{aligned}
$$



KCaM Example





R:

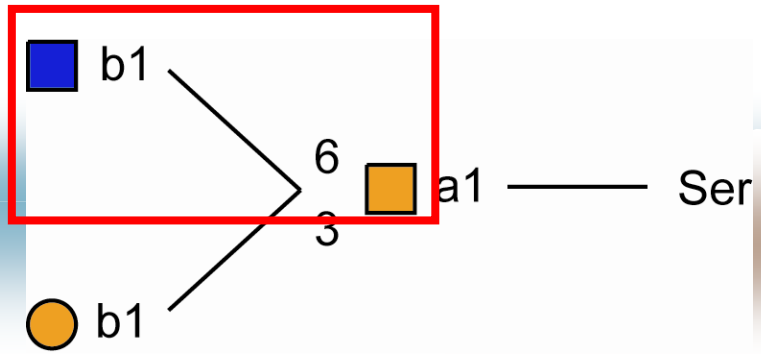
	1	2	3	4	5	6	7	8
A	4	1	0	0	0	0	0	0
B	5	4	1	0	0	0	0	0
C	2	4	3	1	1	0	0	0
D	0	1	3	3	2	1	1	1
E	0	0	1	1	1	1	1	1
F	0	0	1	1	1	1	1	1

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

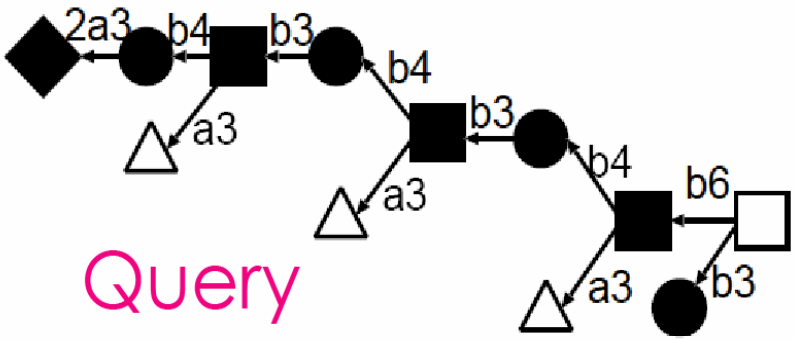
Method

- 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

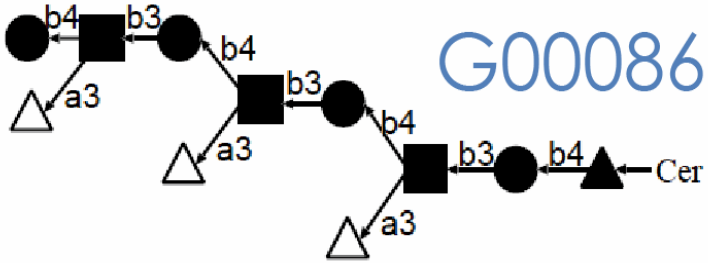


Improved alignments

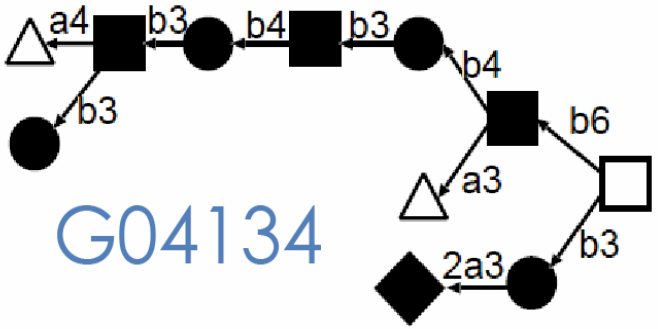
Without Matrix		With Matrix	
Glycan	Score	Glycan	Score
G00086	8.0	G04134 *	5.90797
G00192	8.0	G04072 *	5.54569
G04134 *	7.0	G05073 *	5.20216
G04906 *	7.0	G04906 *	5.09453
G00407 *	6.0	G05305 *	4.99696
G00975	6.0	G04140 *	4.9072



Query



G00086



G04134

Individual Matrix Entries

Aligned Linkage Child	Aligned Linkage Parent	Score
Fuc1, a6GlcNAc	Fuc1, a6GlcNAc	2.45254
GlcNAc1, b4GlcNAc	GlcNAc1, b4GlcNAc	2.37549
Man1, b4GlcNAc	Man1, b4GlcNAc	2.32516
Glc1, b4GlcNAc	Glc1, b4GlcNAc	2.08472
Man1, a4Glc	Man1, a6Glc	2.03222
Man1, a3Glc	Man1, b3Glc	1.9977
Glc1, a2Glc	Glc1, a2Glc	1.99001
Glc1, a3Glc	Glc1, a3Glc	1.98493
GlcNAc1, b6GalNAc	GlcNAc1, b6GalNAc	1.96005

Current glycome informatics

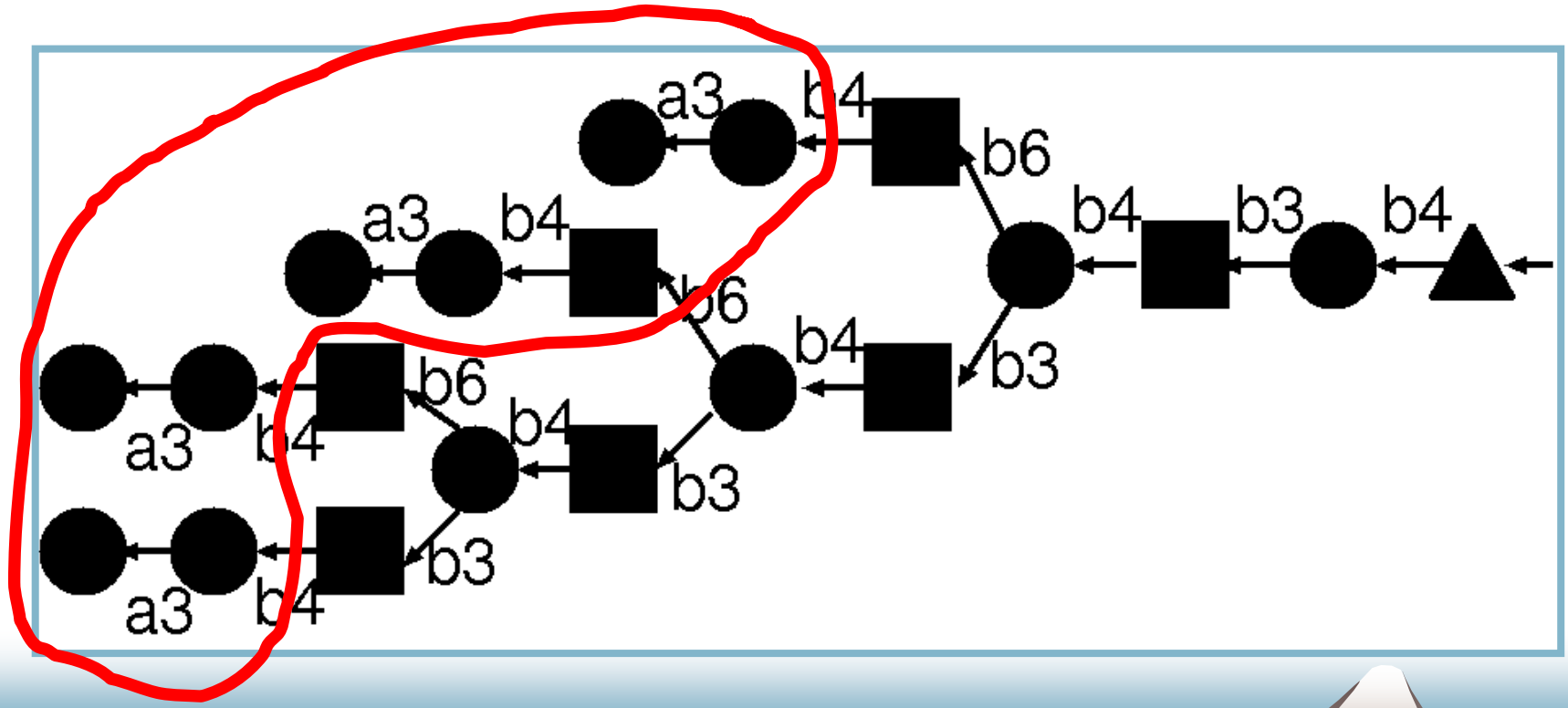
- Automated mass spectrometry annotation
- Computer-theoretic algorithms for tree alignments
- Probabilistic models (mining) for patterns in glycans
- Kernel methods for glycan classification

Mining in Glycome Informatics

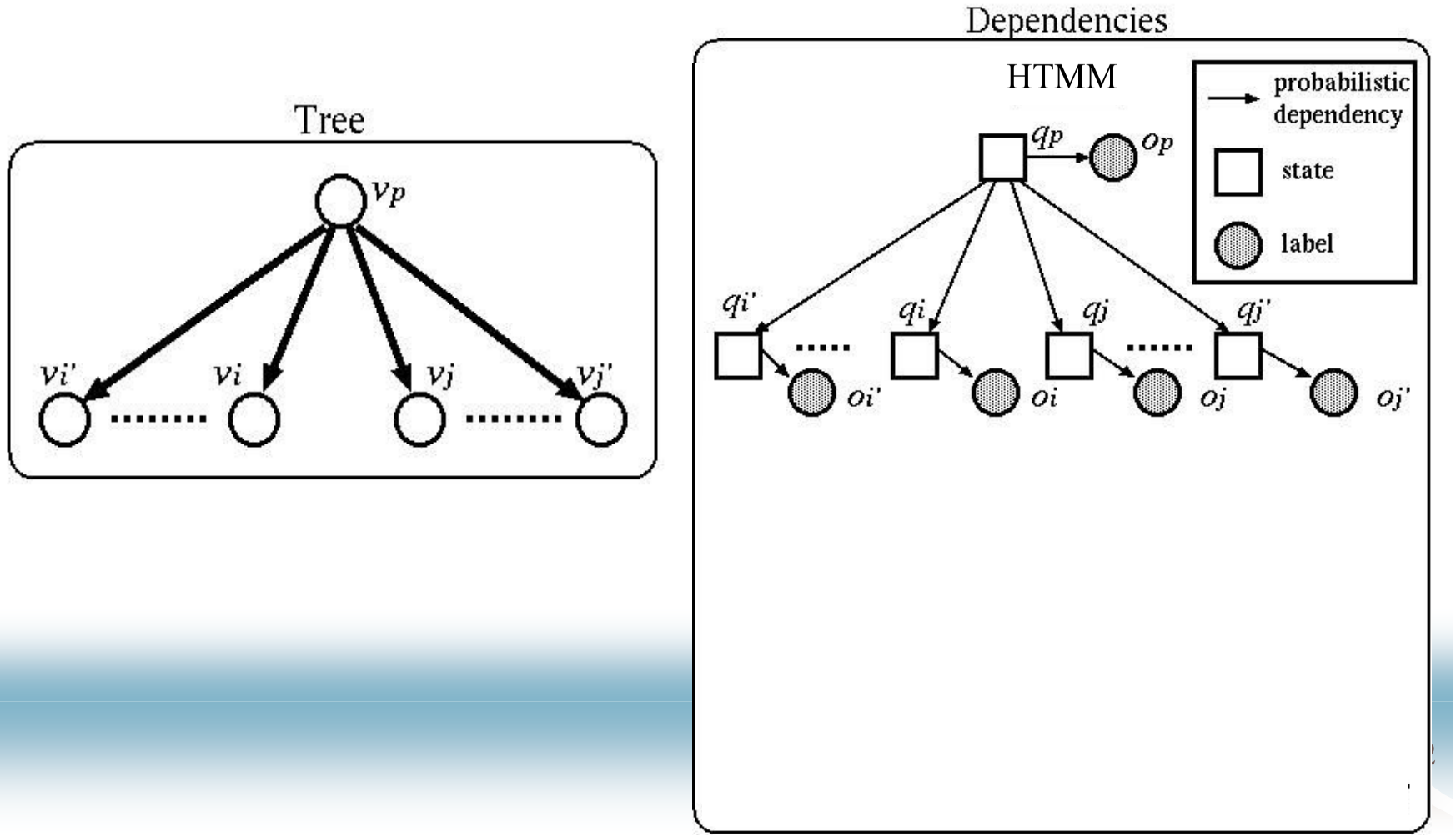
- Probabilistic Models
  - PSTMM, N. Ueda et al, TKDE, 2005
  - Profile PSTMM, K.F. Aoki-Kinoshita et al, ISMB 2006
  - OTMM, Hashimoto et al, KDD 2006

- Previous work on probabilistic trees
  - Hidden Tree Markov Model, HTMM (Diligenti et al., 2003) for image classification

HTMM Cannot Capture Sibling Dependencies!



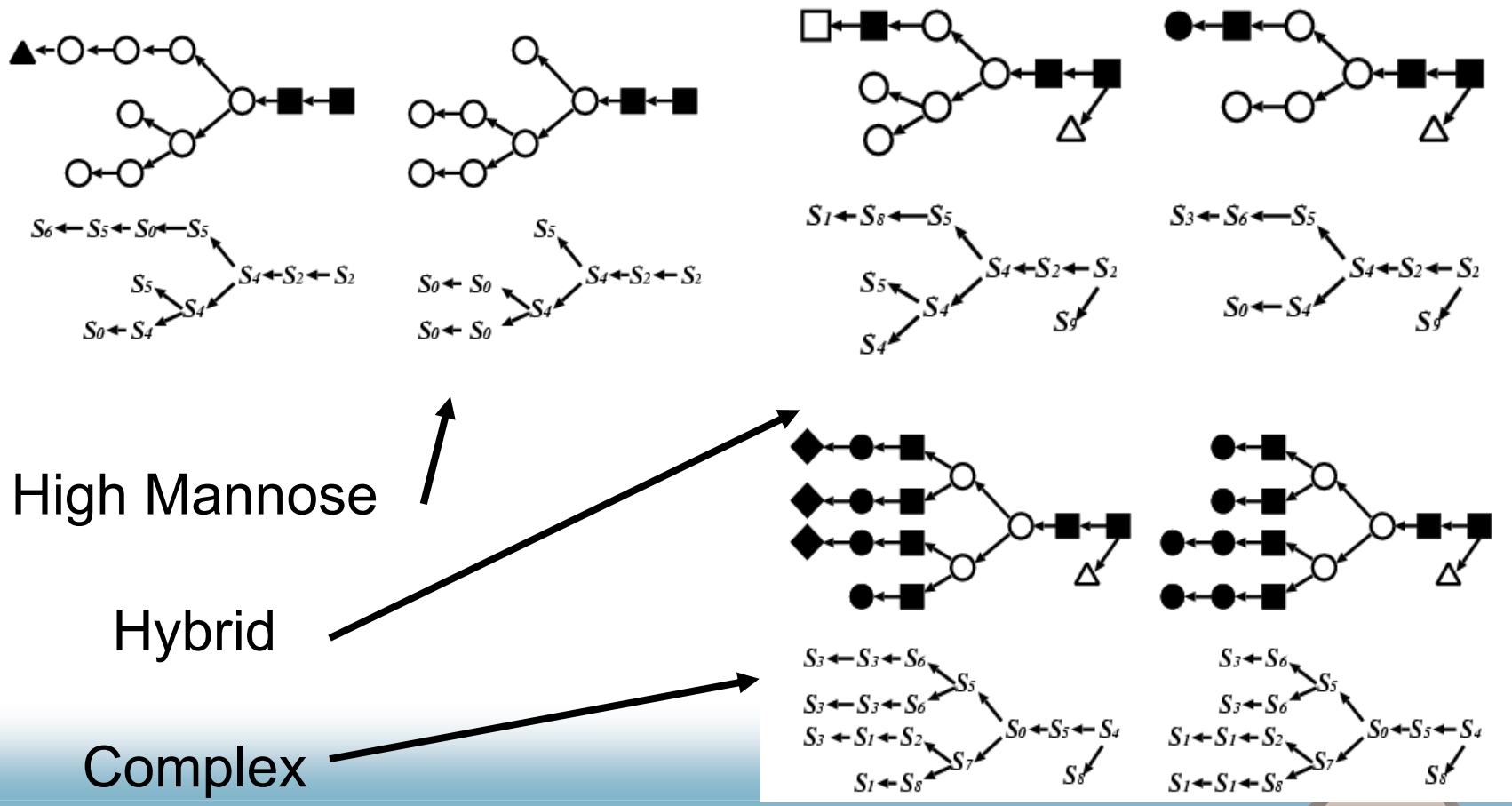
Probabilistic Sibling Tree Markov Model (PSTMM)



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 same class as learned data
- Finding the most likely state transition:
  - To retrieve the learned patterns
  - To apply to multiple tree alignments

Learned Classification

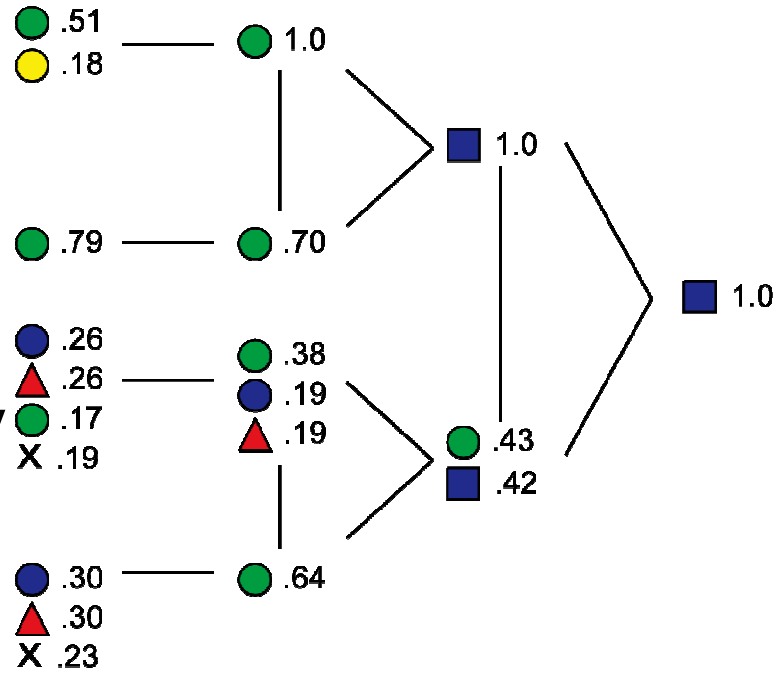


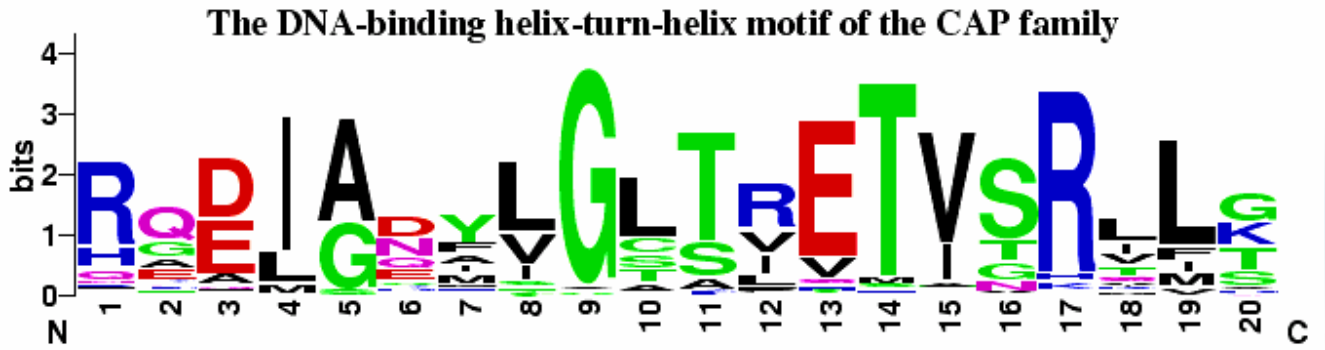
Summary of PSTMM Results

- 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

Profile PSTMM

- 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

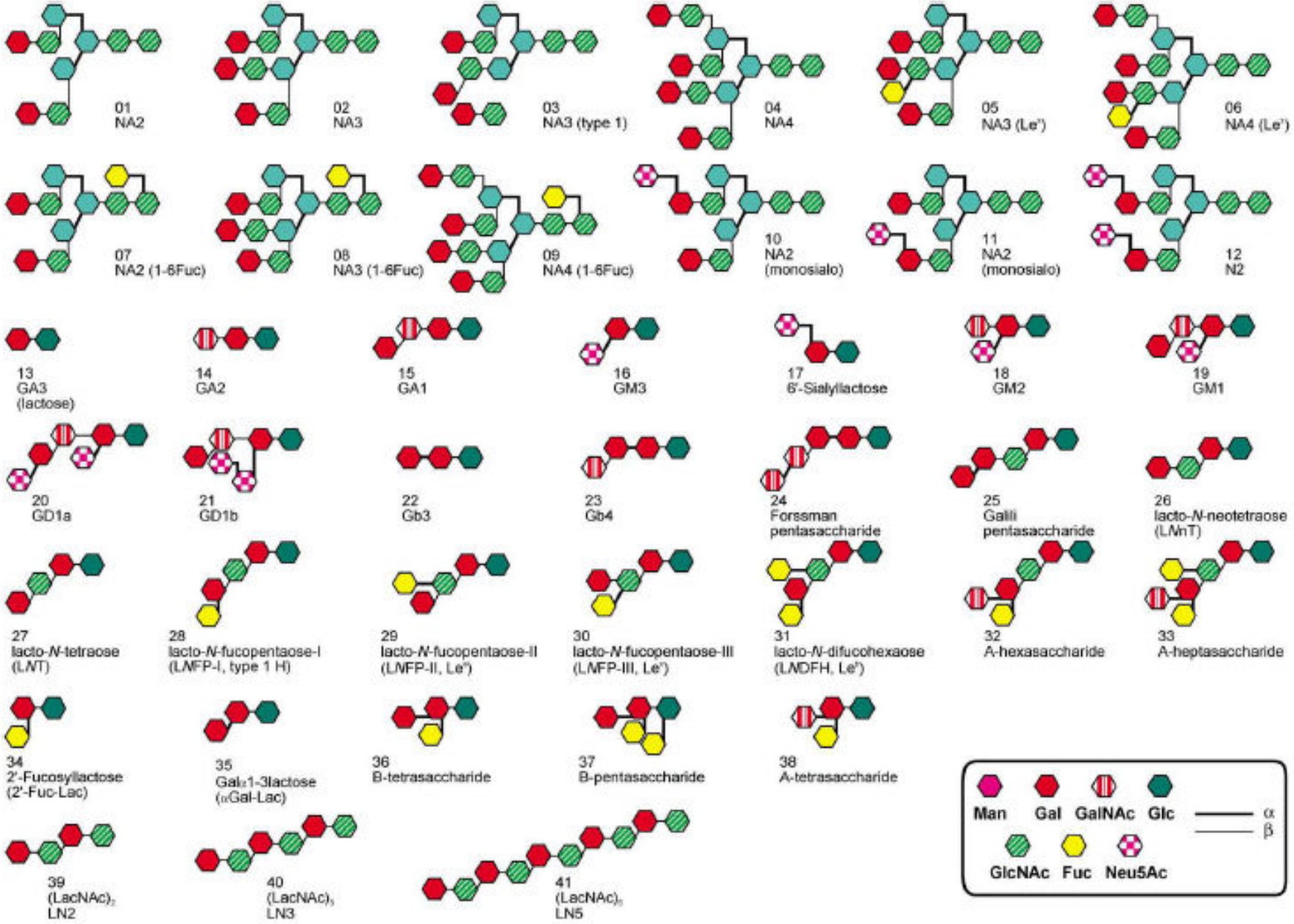
Lectin-glycan experiment

- 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)

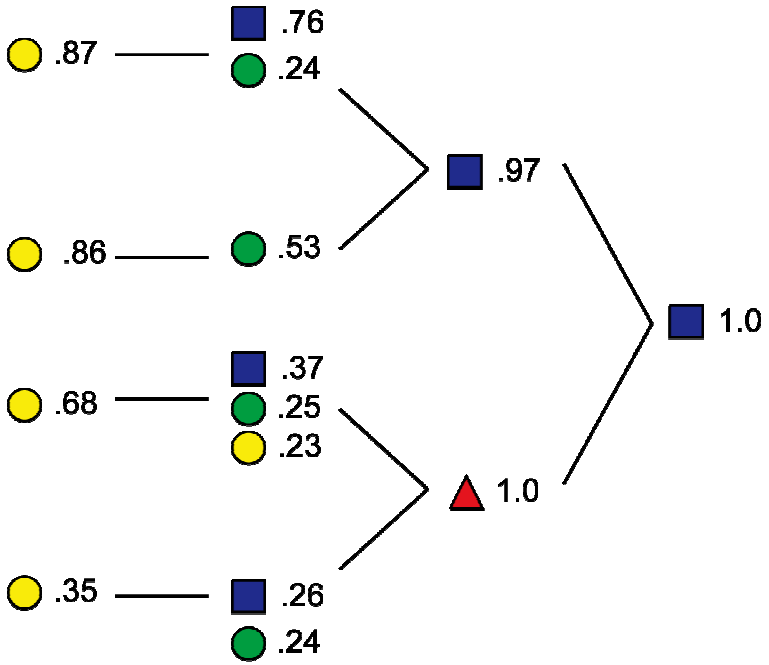
Lectin-glycan experiment

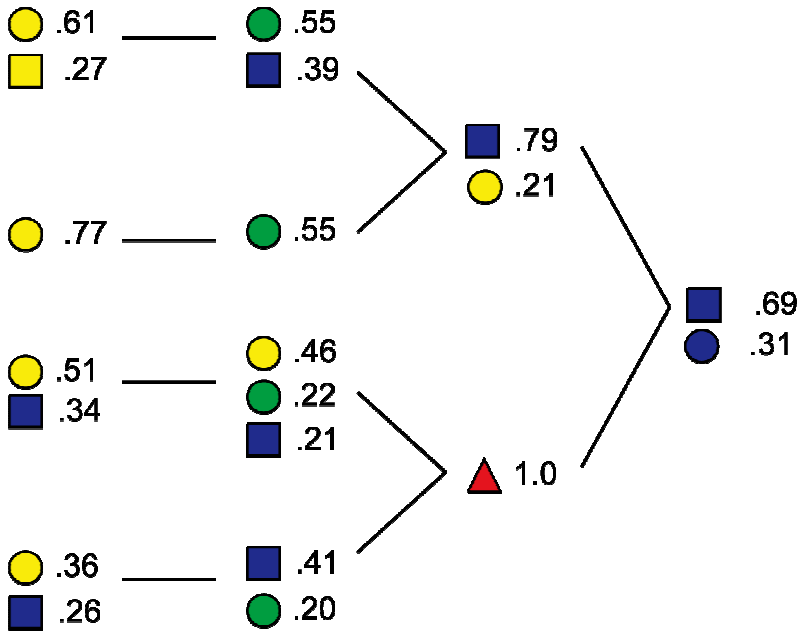
	Gal-3 affinity (weight)	Gal-9N affinity (weight)
NA3	1.28205 (1)	2.6316 (2)
fuc. NA3	1.21951 (1)	2.2222 (2)
NA3 type1	1.08696 (1)	1.6949 (0)
NA4	1.44928 (1)	5.5556 (5)
fuc. NA4	1.40845 (1)	4.3478 (4)
Galili penta.	1.47059 (1)	0.2273 (0)
Forssman penta.	0.16129 (0)	11.111 (11)
A-hexa	1.5873 (1)	3.8462 (3)
LN3	2.85714 (2)	1.2346 (0)
LN5	5.26316 (5)	8.3333 (8)

J. Hirabayashi, et al. Oligosaccharide specificity of galectins: a search by frontal affinity chromatography. Biochim Biophys Acta, 1572(2–3):232–54, 2002.



Lectin-binding glycan profiles





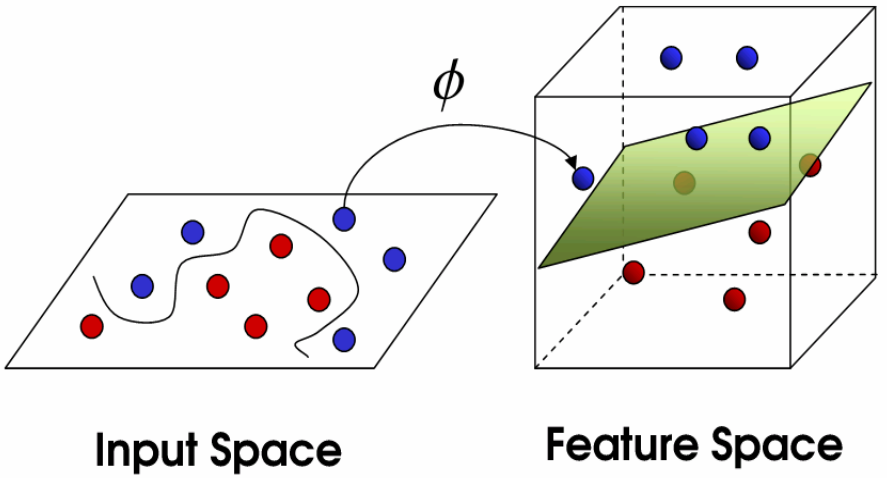
	Gal-3	Gal-9N
Accuracy	.847	.910
Precision	1.0	.918
AUC	.930	.931

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



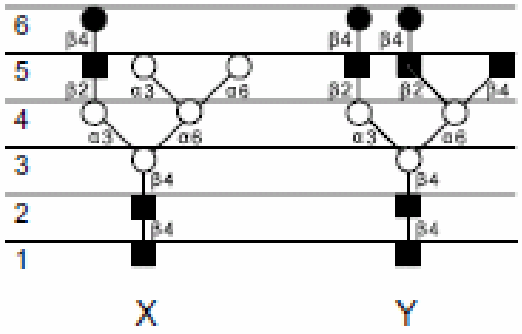
http://www-kairo.csce.kyushu-u.ac.jp/~norikazu/research.en.html

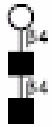
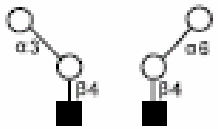
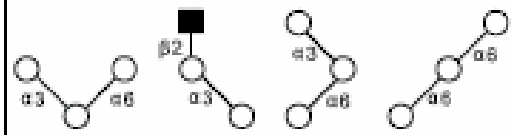
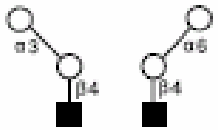
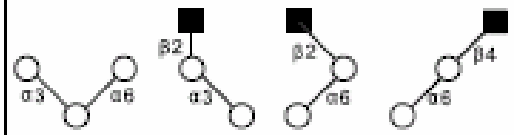
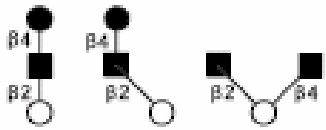
Leukemia-specific features

- Hizukuri et al, Carbohydr. Res. 340, 2270-2278 (2005).
- 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



layer	1	2	3	4
X				
Y				

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, ... x_n)
- Glycans X and Y are compared by the following function:

$$\mathrm{sim}(X, Y) = \sum_{k=1}^{260} w_k x_k y_k, \qquad (2)$$

where w_k is defined as

$$\begin{aligned} w_k &= 1 - \exp(-\alpha h) \quad \text{if } h > 1, \\ w_k &= 1 \qquad\qquad\qquad\quad \text{if } h = 1, \end{aligned} \qquad (3)$$

where h is the layer of the matching substructures and α 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

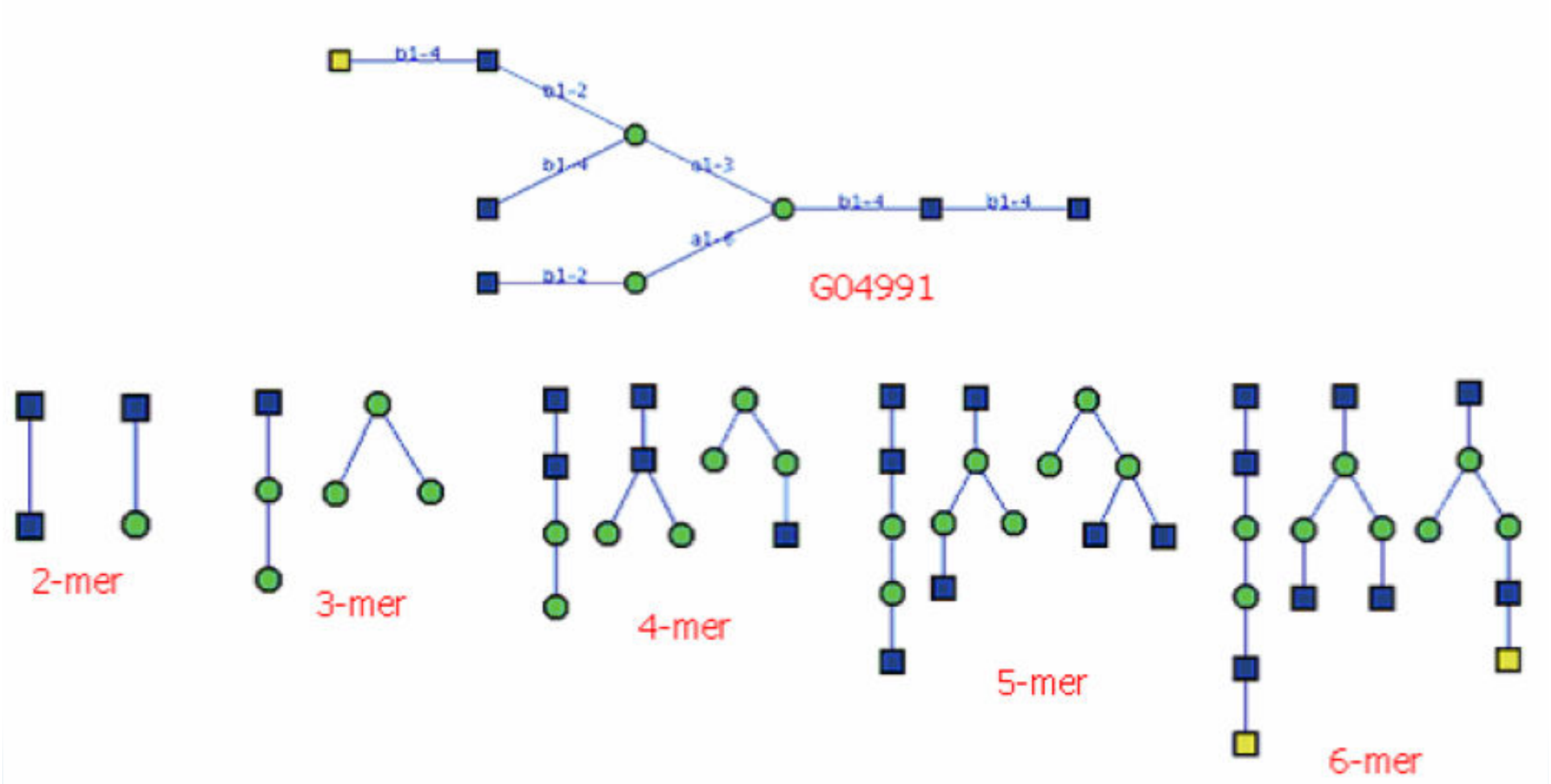
- Supported experimental results

Substructures	Layer	Scores
Leukemic cells		
α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	5	161.2
β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p*	4	159.6
α-D-Neu*p*5Ac-(2→6)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	5	148.8
β-D-Glc*p*NAc-(1→2)-α-D-Man*p*-(1→3)-D-Man*p*	3	78.7
β-D-Glc*p*NAc-(1→2)-α-D-Man*p*-(1→6)-D-Man*p*	3	77.6

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”

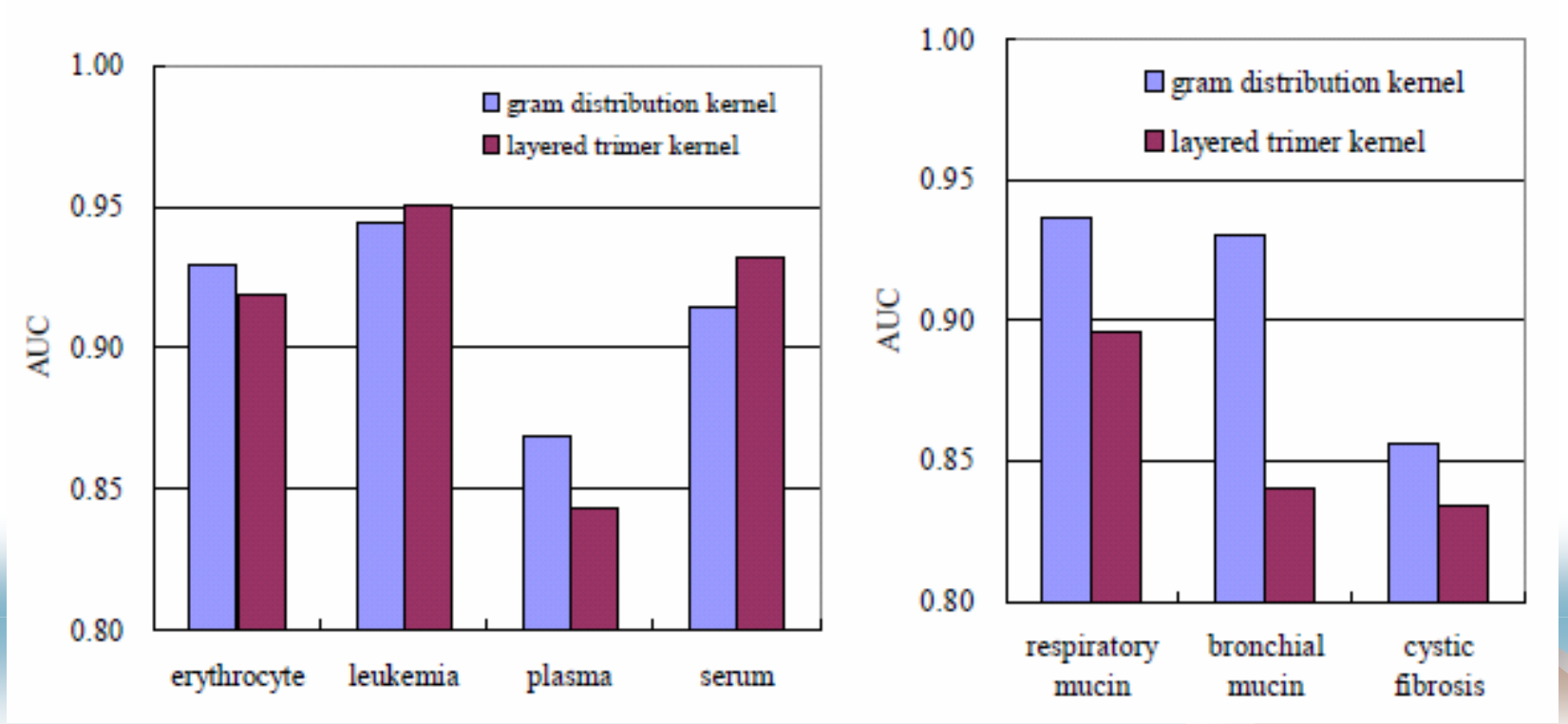
Results: Features extracted

leukemia	
Score	Substructure
226	(leaf)NeuAc2–α3Gal–β4
201	(leaf)NeuAc2–α6Gal–β4
201	(leaf)NeuAc2–α3Gal–β4GlcNAc–β2
200	–Gal–β4GlcNAc–β2Man–α6Man–β4GlcNAc–β4GlcNAc(root)
200	–Gal–β4GlcNAc–β2Man–α3Man–β4GlcNAc–β4GlcNAc(root)

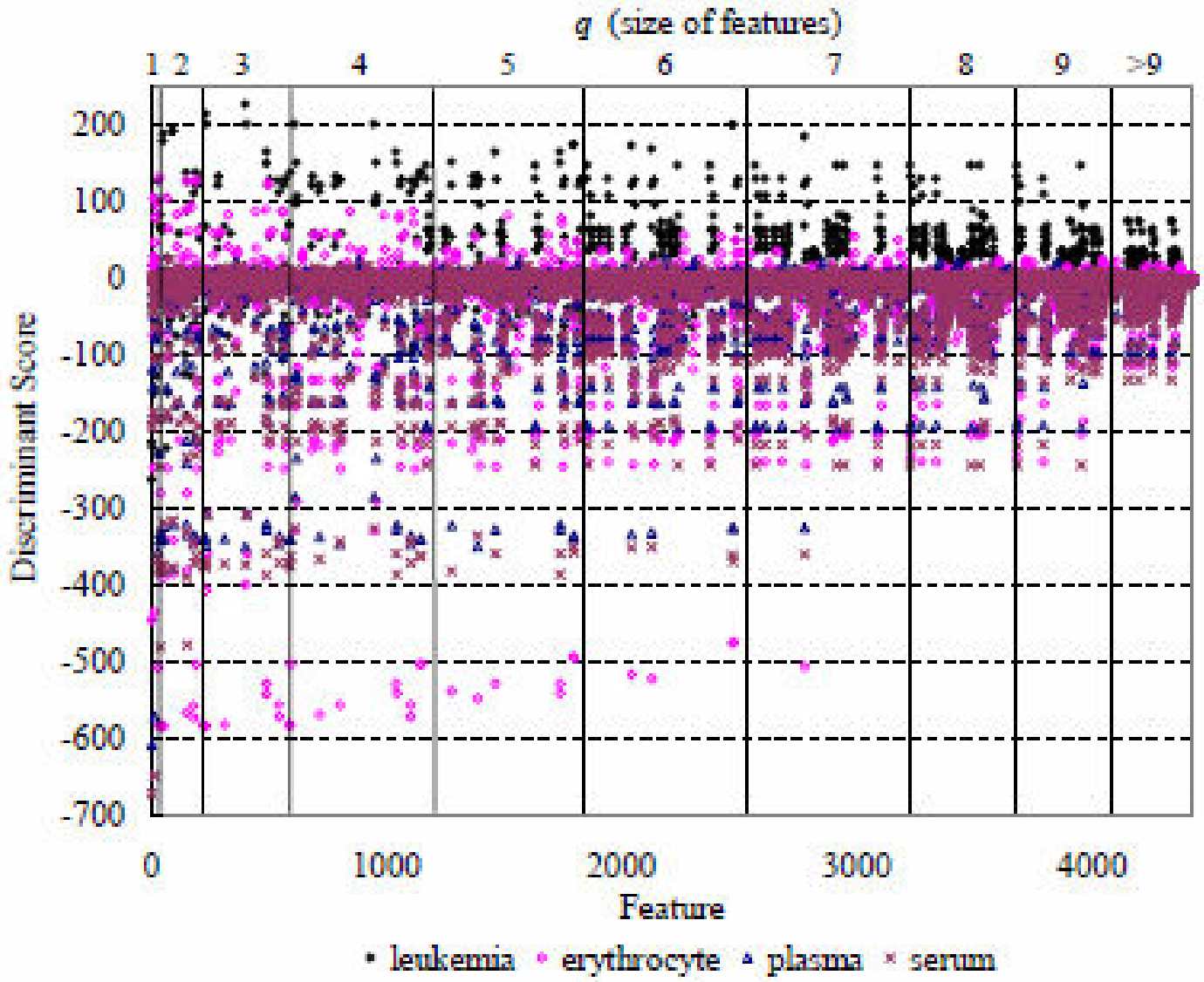
erythrocyte	
Score	Substructure
122	–Gal–β4GlcNAc–β3Gal–β4
86	(leaf)Fuc–α2Gal–β4GlcNAc–β3
86	–GlcNAc–β3Gal–β4Glc–β1(root)
82	–Gal–β4GlcNAc–β3Gal–β4GlcNAc–β3
81	(leaf)Fuc–α2Gal–β4GlcNAc–β3Gal–β4

Results: performance

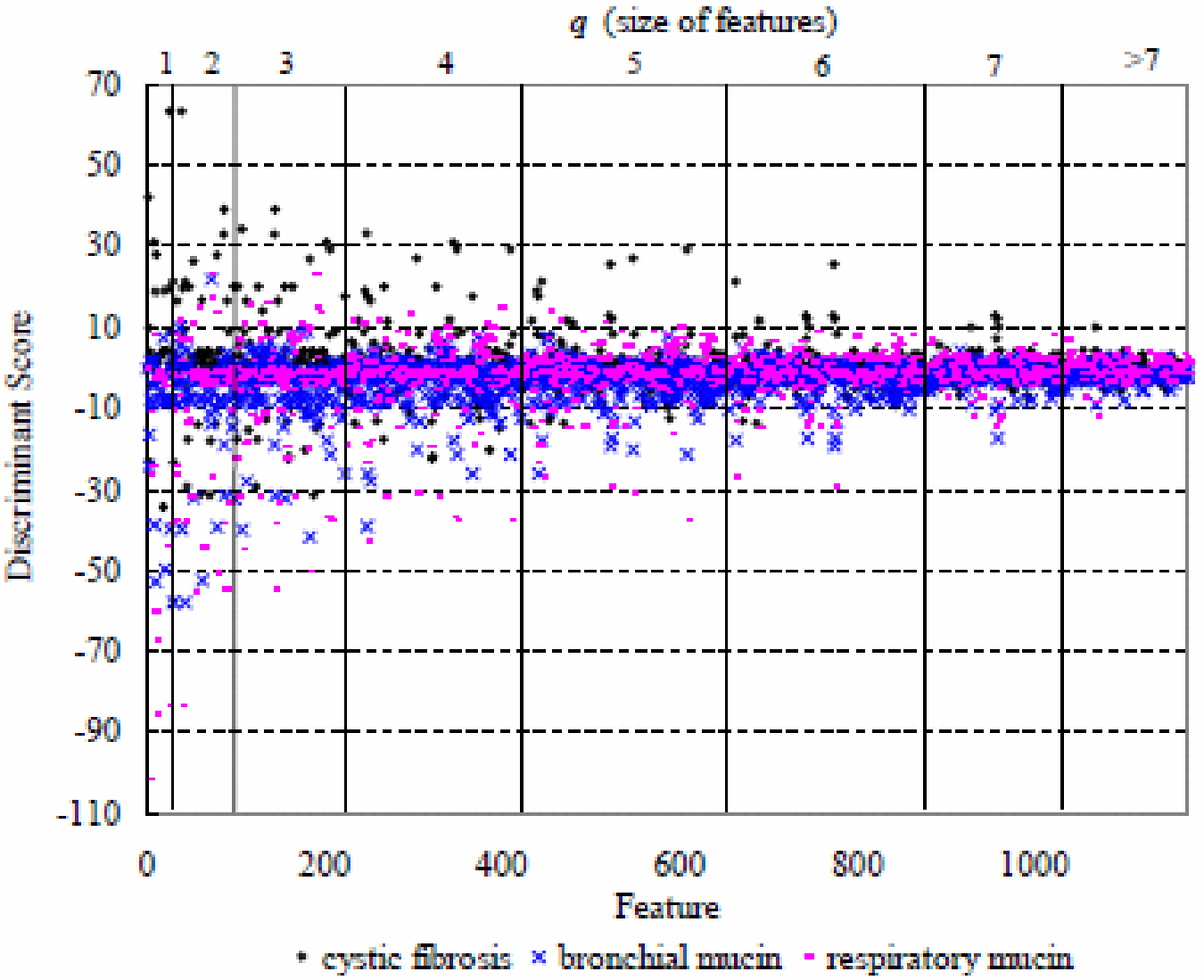
- Gram distribution vs. Leukemia kernel (layered trimer kernel)



Results: marker size



Results: marker size



Systems approach to unveiling structure-function relationship

Unknown structural space for glycan structure

Glycan synthesis is 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 = $E^n \times 2^n_{(anomer)} \times 2^n_{(conf)} \times (4^{n-1})$

Monosaccharid	1	4
Disaccharide	2	256
Trisaccharide	3	27,648
Tetrasaccharide	4	4,194,304
Pentasaccharide	5	819,200,00
Hexasaccharide	6	195,689,447,424

Which glycan structures really exist in certain species ? What do the databases say ?

Occurrence of monosaccharide residues (CarbBank nomenclature)

Mammalian: 5339
Human : 2128

Total number of different residues

Mammalian : 86
Human : 83

Ź	Monosaccharide name	mammalian #	Mammalian [%]	human #	human [%]	
1	B-D-GLCPNAC	7319	26,1%	4705	26,69%	
2	B-D-GALP	6389	22,8%	4178	23,70%	
3	A-D-MANP	3659	13,1%	2073	11,76%	
4	A-D-NEUP5AC	2101	7,5%	1465	8,31%	
5	A-L-FUCP	1971	7,0%	1461	8,29%	
6	B-D-MANP	1486	5,3%	900	5,10%	
7	D-GLCNAC	675	2,4%	403	2,29%	
8	D-GLCNAC-OL	598	2,1%	399	2,26%	
9	D-GALNAC-OL	511	1,8%	355	2,01%	88.4%
10	B-D-GLCP	423	1,5%	244	1,38%	
11	B-D-GALPNAC	431	1,5%	230	1,30%	
12	*SULFATE*	450	1,6%	198	1,12%	
13	A-D-GALPNAC	248	0,9%	171	0,97%	
14	D-GLC	197	0,7%	151	0,86%	
15	A-D-GALP	287	1,0%	103	0,58%	
16	D-GALNAC	116	0,4%	91	0,52%	
17	A-D-GLCP	161	0,6%	68	0,39%	
18	B-D-GLCPA	94	0,3%	54	0,31%	
19	D-GAL	56	0,2%	37	0,21%	97.5%
20	D-GLC-OL	37	0,1%	34	0,19%	
21	A-D-GLCPNAC	89	0,3%	31	0,18%	
22	D-GAL-OL	38	0,1%	27	0,15%	
23	D-GLCPNAC	37	0,1%	17	0,10%	
24	A-D-NEUP5GC	132	0,5%	16	0,09%	
25	B-D-XYLP	23	0,1%	13	0,07%	
26	D-GALP	22	0,1%	12	0,07%	
27	A-L-4-EN-THRHEXPA	40	0,1%	12	0,07%	
28	?-D-GALPNAC	15	0,1%	11	0,06%	
29	*P*	20	0,1%	11	0,06%	99.1%
30	D-2,5-ANHYDRO-MAN-OL	13	0,0%	9	0,05%	

10

30

Occurrence of disaccharide residues (CarbBank nomenclature)

Human :2128

Total number of different Disaccharide

Human : 171

once : 65

twice : 20

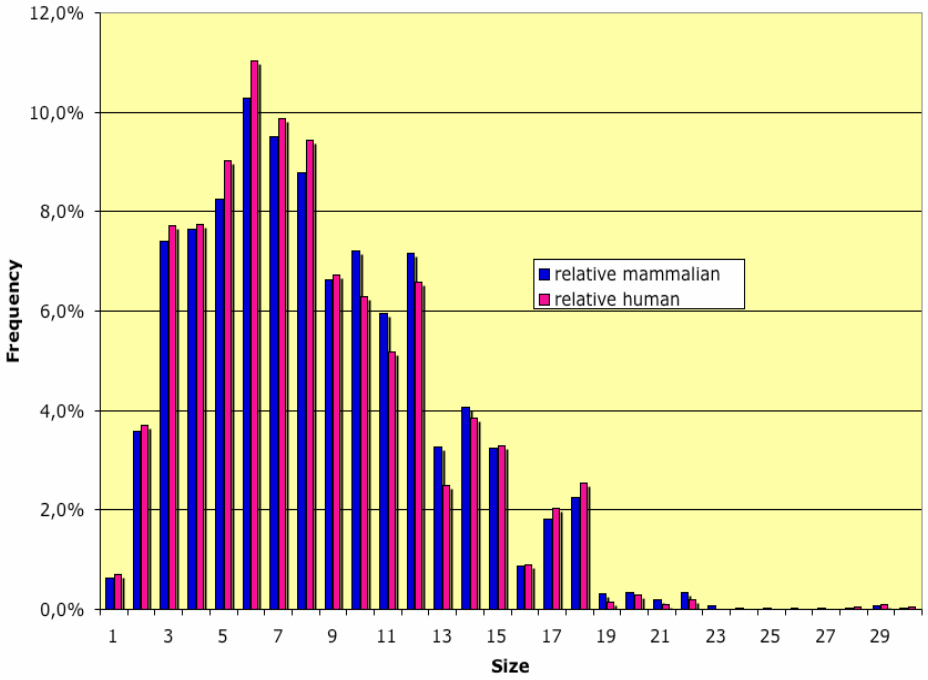
Three Times: 10

	Parent	from	to	Child	#	
1	B-D-GLCP-2NAC	4	1	B-D-GALP	2837	
2	A-D-MANP	2	1	B-D-GLCP-2NAC	1382	
3	B-D-GALP	3	1	B-D-GLCP-2NAC	860	
4	B-D-MANP	6	1	A-D-MANP	776	
5	B-D-MANP	3	1	A-D-MANP	771	
6	B-D-GALP	3	2	A-D-NEUP-5AC	742	
7	B-D-GLCP-2NAC	4	1	B-D-MANP	732	
8	B-D-GALP	6	2	A-D-NEUP-5AC	467	
9	B-D-GALP	2	1	A-L-FUCP	436	71.7%
10	B-D-GLCP-2NAC	3	1	A-L-FUCP	418	
11	A-D-MANP	4	1	B-D-GLCP-2NAC	340	
12	B-D-GLCP-2NAC	3	1	B-D-GALP	300	
13	A-D-MANP	6	1	B-D-GLCP-2NAC	255	
14	B-D-GLCP	4	1	B-D-GALP	219	
15	B-D-GALP	6	1	B-D-GLCP-2NAC	186	
16	B-D-GLCP-2NAC	4	1	B-D-GLCP-2NAC	175	
17	A-D-MANP	2	1	A-D-MANP	156	
18	B-D-GALP	3	1	A-D-GALP-2NAC	119	
19	B-D-GLCP-2NAC	4	1	A-L-FUCP	117	89.9%
20	B-D-GLCP-2NAC	4	1	B-D-GALP-2NAC	110	
21	A-D-MANP	3	1	A-D-MANP	92	
22	A-D-MANP	6	1	A-D-MANP	88	
23	B-D-GLCP-2NAC	6	1	A-L-FUCP	86	
24	B-D-MANP	4	1	B-D-GLCP-2NAC	78	
25	B-D-GALP	3	1	A-D-GALP	68	
26	B-D-GALP	4	1	B-D-GALP-2NAC	62	
27	B-D-GALP-2NAC	3	1	B-D-GALP	45	
28	A-D-GALP-2NAC	3	1	B-D-GALP	39	95.5%
29	A-D-NEUP-5AC	8	2	A-D-NEUP-5AC	31	

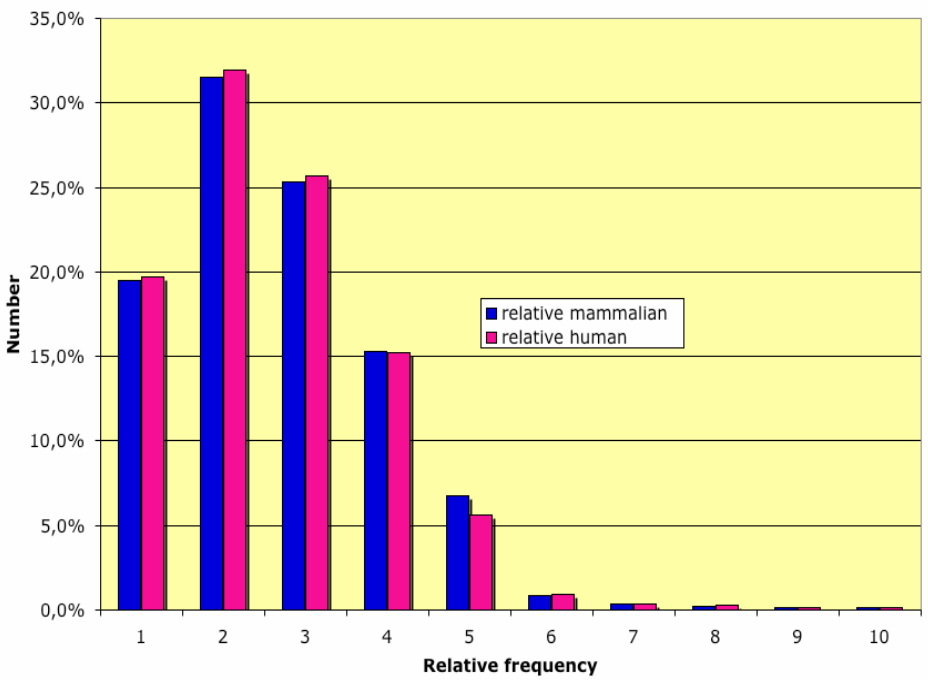
30

Topologies of Glycans

Size of Glycan (Residues)



Number of Branching points



Mathematical Modelling to explore the structural space of glycan using Information from carbohydrate active enzymes

A Mathematical Model of N-Linked Glycosylation

Frederick J. Krambeck, Michael J. Betenbaugh; Biotechnol Bioeng. 2005 Dec 20;92(6):711-28.

Enzyme reaction rule tables to model reaction networks:

Parameters: spatial distribution of enzymes, transport, reaction kinetcs, donor concentrations.

Enzymes included	
Abbreviations	**EC No.**
ManI	**3.2.1.113**
ManII	**3.2.1.114**
FucT	**2.4.1.68**
GnTI	**2.4.1.101**
GnTII	**2.4.1.143**
GnTIII	**2.4.1.144**
GnTIV	**2.4.1.145**
GnTV	**2.4.1.155**
GnTE	**2.4.1.149**
GalT	**2.4.1.38**
SiaT	**2.4.99.6**

Glycoform description scheme	
Man	Number of mannose residues
Fuc	Number of fucose residues.
Gnb	Number of bisecting GlcNAc residues
Gal	Number of galactose residues
Sia	Number of sialic acid (NeuAc) residues
Br1	Extension level of branch 1.
Br2	Extension level of branch 2.
Br3	Extension level of branch 3.
Br4	Extension level of branch 4.

The full model generates 7565 N-glycan structures in a network of 22,871 reactions

Distribution of carbohydrate chains in PDB
(Release September 2004)

	N-glycan		O-glycans		ligands		total	
Chain length	#	%	#	%	#	%	#	%
1	1977	58.5	555	91.4	2093	59.0	4625	61.4
2	693	20.5	29	4.8	812	22.9	1534	20.4
3	310	9.2	10	1,7	329	9.3	649	8.6
4	83	2.5	2	0.3	149	4.2	234	3.1
5	98	2.9	5	0.8	83	2.3	186	2.5
6	98	2.9	2	0.3	42	1.2	142	1.9
7	55	1.7	4	0.7	15	0.4	74	1.0
8	37	1.1	0	0	15	0.4	52	0.7
total	3381		607		3550		7538	

Frontiers in Glycomics: Bioinformatics and Biomarkers in Disease September 11-13, 2006

"*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-Kinoshita
- kkiyoko@t.soka.ac.jp