Functional annotation of sugar databases

Frederique Lisacek\textsuperscript{1}, Christine Hoogland\textsuperscript{1}, Elaine Mullen\textsuperscript{3}, Lynette Hirschman\textsuperscript{2}, Nicolle Packer\textsuperscript{4}

\textsuperscript{1}Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

\textsuperscript{2}The MITRE Corporation, Mass., USA

\textsuperscript{3}Biomolecular Frontiers Research Centre, Faculty of Science, Macquarie University, Sydney, Australia
GlycoSuiteDB

- N- and O-linked linear and branched glycan structures
  - linkage, anomeric configuration, mass and composition
- biological source information
  - native and recombinant sources
  - tissue/cell type, cell line, strain and disease state
  - protein name (UniProtKB)
  - site of attachment
- methods used
- literature references (PubMed)
**GlycoSuiteDB entry**

*entry: 1187-1149*

- **NeuAc** a2 → 3 **Gal** b1 → 4 **GlcNAc** b1 → 2 **Man** a1
  - **Man** b1 → 4 **GlcNAc** b1 → 4 **GlcNAc**

**Species**  
Homo sapiens (HUMAN)

**Class**  
MAMMALIA

**Source**  
UROGENITAL SYSTEM, EXCRETION, URINE

**Source notes**  
BIOLOGICALLY ACTIVE URINARY hCG PROVIDED BY A NUMBER OF INSTITUTES.

**Attached to**  
CHORIOGNADOTROPIN ETA CHAIN (swiss-prot entry PD1233); amino-acid ASN-33

**Linkage**  
N-LINKED

**Glycosylation sites**  

**Identified by methods**  
PROTON NMR

**References**  
Weisshaar (1991) Glycobiology 1: 393-404

**Glycan structure**  
NeuAc(a2-3)Gal(b1-4)GlcNAc(b1-2)Man(a1-3)[NeuAc(a2-3)Gal(b1-4)GlcNAc(b1-2)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc

**Mass**  
2222.7830 Da (monoisotopic), 2224.0232 Da (avg), total residues: 11

**Composition**  
Hex5HexNAc4NeuAc2

**Release date**  
04-NOV-00 (last updated 04-DEC-00)
Basis of sugar annotation

- Methods for glycan structure determination: 57 listed in GlycoSuite DB, e.g., NMR, methylation, MS, HPLC, etc.
  - Scalable reliability
  - Potential for evidence tags
- Further insight with MS/MS
- Possible scenarios of synthesis:
  - Upon knowledge of enzymes (in genomes, in expression data)
  - Upon more or less established branching rules
MS/MS

Sugar 1 → MS → Sugar 2 → ... → Sugar n

MS/MS

Fragment ions 1 → MS/MS → Fragment ions 2 → MS/MS → Fragment ions 3 → ... → Fragment ions i → MS/MS → Fragment ions j → ... → Fragment ions p

Sugar 2 ▶ Sugar 1

... ▶ Sugar n

© 2009 SIB

Swiss Institute of Bioinformatics
Enzymatic constraints
Recent addition on ExPASy

About This Database

The Pathogen Sugar-Binding Database was developed under MITRE internal funding, as part of a project to investigate use of floating glycoprotein films to capture pathogens. This database provides a list of known carbohydrate sequences to which pathogenic organisms specifically adhere. The data were compiled through an exhaustive search of literature published over the past 30 years by glycobiologists, microbiologists, and medical histologists.

The database allows users to search for bacteria, toxins, and viruses that bind to a particular sequence of sugars at the non-reducing terminus of an oligosaccharide. Alternatively, one can learn which glycans are ligands of adhesins (lectins) expressed by a particular species. Abstracts of primary literature are easily accessed, and a list of citations can be printed from search results.
Host-pathogen interactions via sugars

Mammalian host

Lectin repertoire

Glycosphingolipid

Membrane glycoprotein

Extracellular

Cell membrane

Intracellular

Nature 446, 1023-1029, 2007

Biocuration 2010

Swiss Institute of Bioinformatics
SugarBind

Known epitopes

Lectin repertoire

+ references
### Example

<table>
<thead>
<tr>
<th>Pathogen or Toxin</th>
<th>Fimbria / Pilus</th>
<th>Lectin / Adhesin</th>
<th>Gene Name</th>
<th>? Carbohydrate or Ligand</th>
<th>Pub Year</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IIIL</td>
<td>lecB</td>
<td>Fuc</td>
<td></td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IIIL</td>
<td>lecB</td>
<td>Fuc(a1-4)[Gal(b1-3)]GlcNAc(b1-3)Gal(b1-4)Glc</td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IIIL</td>
<td>lecB</td>
<td>Fuc(a1-4)[Neu(a2-3)Gal(b1-3)]GlcNAc(b1-3)Gal(b1-4)Glc</td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IIIL</td>
<td>lecB</td>
<td>Gal(b1-3)[Fuc(a1-4)]GlcNAc(b1-3)Gal(b1-4)Glc</td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IIIL</td>
<td>lecB</td>
<td>NeuAc(a2-3)Gal(b1-3)[Fuc(a1-4)]GlcNAc(b1-3)Gal(b1-4)Glc</td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IL</td>
<td>lecA</td>
<td>Gal</td>
<td></td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IL</td>
<td>lecA</td>
<td>Gal(a1-3)Gal</td>
<td></td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IL</td>
<td>lecA</td>
<td>Gal(a1-4)Gal</td>
<td></td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>PA-IL</td>
<td>lecA</td>
<td>Gal(a1-6)Gal</td>
<td></td>
<td>2004</td>
<td>Microbes and Infection, 2004 Feb; vol. 6 no. 2: 221-8</td>
</tr>
</tbody>
</table>
Open consortium

- **N. Packer et al**, Biomolecular Frontiers Research Centre of Macquarie University, Sydney, Australia
- **P. Rudd et al**, Dublin-Oxford Glycobiology Lab of the Conway Institute in Dublin, Ireland
- **N. Karlsson et al**, Medical Biochemistry Dpt of University of Gothenburg in Sweden
- **E. Mullen et al**, MITRE Corporation in Massachusetts, USA
- The more the merrier ...
Roadmap (1)

- **Step 1**: Define the basis of extensive linking of glycomics data (*standards* need refinement and reinforcement)
- **Step 2**: Link data production to structural determination (*manual curation* of MS1/MS2 data is necessary for the accurate determination of structural data)
- **Step 3**: Link sugar structures with protein attachment site (*existing databases* play central role for collecting data on structures released from specific proteins/tissues)
Roadmap (2)

• Step 4: Link mass spectrometry data to the expression of glycosyltransferases or glycosidases (*reverse proteomics*)

• Step 5: Link glycosylation information to cellular function (understanding of *functional role* of sugars) [strong human input, “manual” data curation]

*In brief*: knowledge of structure/function of glycosylation of proteins to be determined in much the same way as knowledge structure/function of proteins integrated into UniProtKB
© 2009 SIB

Influence of Geneva...
Immediate concerns

• **Challenge 1:** Strengthen functional annotation in host-pathogen interactions by linking epitopes to full sugar structures on glycolipids and glycoproteins

• **Challenge 2:** Automate MS data processing for feeding MS and MS/MS into databases

• **Challenge 3:** Make sugar analysis easier using bioinformatic tools!
Acknowledgements

- SIB colleagues
- PIG members
- Consortium members

Thank you!