GlycoDigit: An alpha-numeric representation system to visualize, store and compare glycan structure data.

Faraaz NK Yusufi[1], Wonjun Park[1], May May Lee[1], Dong-Yup Lee[1,2]

[1] Bioprocessing Technology Institute, Biomedical Sciences Institutes, 20 Biopolis Way, #06-01 Centros, Singapore 138668
[2] Department of Chemical & Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576
Contact: Lee_Dong_Yup@bti.a-star.edu.sg or cheld@nus.edu.sg; Tel: +65-6478-8900

Abstract

Glycans are complex chains of monosaccharides that are attached to proteins during post-translational modification. Secreted recombinant proteins produced in mammalian cell cultures display a wide range of glycan structures that are attached to them. Increasing amounts of glycome data are being generated by wet-lab experiments, aiming to control the glycoform profile of recombinant proteins. At the same time, computational scientists are building *in-silico* experiments to simulate the complex interaction pathways involved in the glycosylation process. Both of these approaches to studying glycosylation would benefit greatly from a compact notation for representing glycan structures that can be easily stored and interpreted by computers.

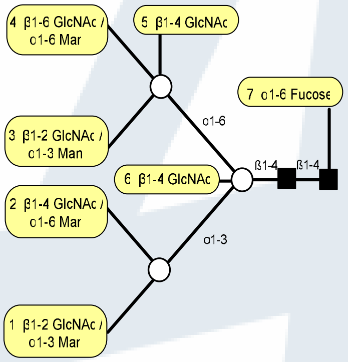
We propose a fixed-length alpha-numeric code for representing N-linked glycan structures commonly found in secreted glycoproteins from mammalian cell cultures. This code, GlycoDigit, uses a pre-assigned alpha-numeric index to represent the monosaccharides attached in different branches to the core glycan structure. The branch-centric representation allows biologists to visualize the structure while the numerical nature of the code makes it machine readable. In addition, this code can be easily extended to define a difference operator, and to develop algorithms that can be used to measure the similarity between structures and generate the necessary reaction steps to convert one structure to another.

Representing N-linked Glycan Structures

N-linked glycosylation occurs in all eukaryotic cells with the glycans sharing a common pentasaccharide core structure. Chains of glycans can attach at seven possible sites on this core structure.

Based on these seven possible linkage sites, GlycoDigit uses seven digit-letter pairs to represent glycan structures.

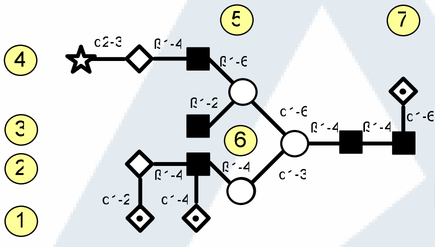
The digit portion of each pair corresponds to the number of monosaccharides attached at that branch while the letter serves as an index to a table containing additional information about the type of linkage and the specific sugar molecule added.



■ GlcNAc ○ Mannose ◈ Fucose ◇ Galactose ☆ NeuNAc

There are three types of N-Glycan structures: High-Mannose, Complex and Hybrid Type. GlycoDigit can be used to represent all three types of structures.

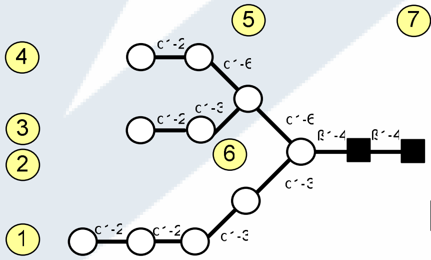
Complex Glycan



GlycoDigit Code

0x 3g 1a 3a 0x 0x 2e

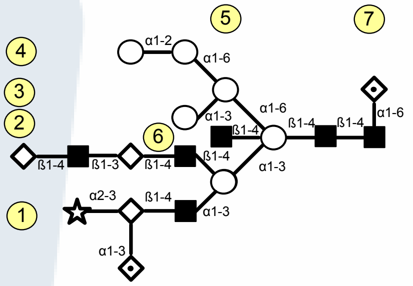
High-Mannose Glycan



GlycoDigit Code

Ca 0x Ba Ba 0x 0x 0x

Hybrid Glycan

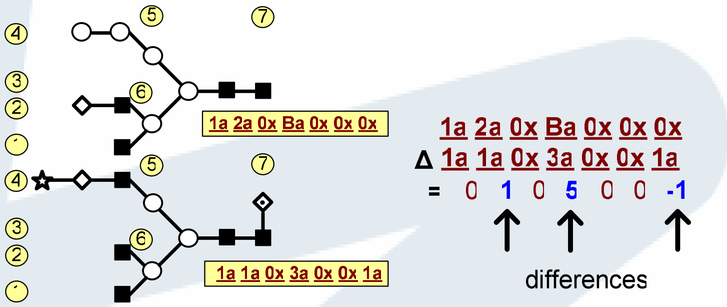


GlycoDigit Code

3b 4a Aa Ba 0x 1a 2a

Differences Between Glycan Structures

The numerical nature of GlycoDigit makes it convenient to define a difference operator Δ which allows for easy comparison of different glycan structures.



1a 2a 0x Ba 0x 0x 0x

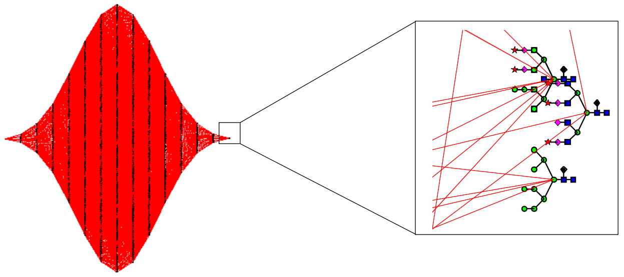
1a 1a 0x 3a 0x 0x 1a

```
  1a 2a 0x Ba 0x 0x 0x
Δ 1a 1a 0x 3a 0x 0x 1a
=  0  1  0  5  0  0 -1
```

The non-zero positions in the resulting code from the difference operator reveal which branches the two structures differ in. The result code can also be used to calculate the number of reaction steps necessary to convert one structure to another. Adding the absolute values of the digits in the difference code reveals the number of reactions needed, in this case 7 reactions would be needed to convert the first structure to the second.

Visualizing Glycan Networks

The glycosylation reaction network can be thought of as a graph with the nodes representing glycan structures and edges showing possible enzymatic reactions. A single glycan structure can act as a substrate to multiple reactions and also be the end product of several reactions, thus creating a highly branched network. An adjacency matrix for approximately 5000 glycans was created to record the edges in the network and populated using the difference operator Δ. In order to visualize the glycosylation network, glycans were arranged from the basic core structure and sugar residues were added until the structure was fully sialylated. Glycans were classified into groups based on the number of reaction steps that separated each glycan from the core structure.



Conclusion and Future Work

We presented an alpha-numeric code, GlycoDigit, which is based on a pre-defined branching structure of N-linked glycans that are commonly found in most mammalian cells. Compared to other standard text representations for glycans, GlycoDigit is much shorter and more intuitive as it focuses on branches instead of previous methods that describe individual monosaccharide units. This code is adaptable and interoperable, which allows us to incorporate it into a laboratory glyco-information management system. Web based tools to graphically represent and compare glycan structures are currently under development to demonstrate the usefulness of this representation.

18th International Conference on Genome Informatics, GIW, December 2007, Singapore