Automatic literature mining tool to extract glycosylation information from literature

Rahi Navelkar[1], Jeet Vora[1] Karen Ross[2], Catherine Hayes[3], Frederique Lisacek[3] ,Peng Su[4] ,Vijay Shanker[4]

[1]Department of Biochemistry & Molecular Medicine, The George Washington School of Medicine and Health Sciences, Washington, DC 20052, USA.,
[2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC 20007, USA.,
[3] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.,
[4]University of Delaware, Newark, DE.

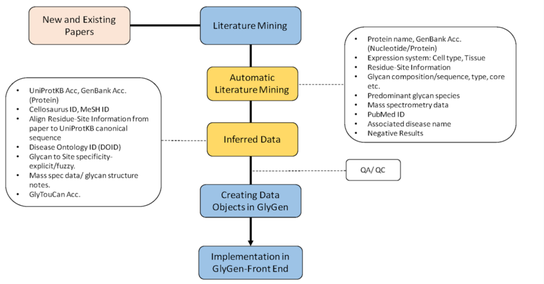
Presenter: Jeet Vora , Catherine Hayes

Abstract

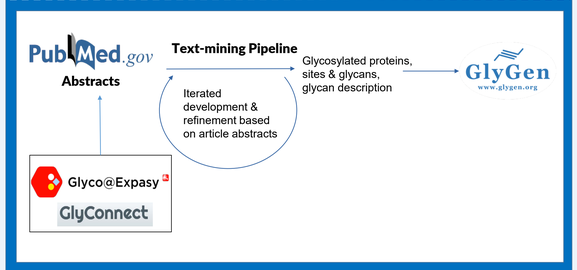
Glycosylation is one of the most common and complex post-translational phenomena which impacts several key biological processes making it vital to study it with regard to human health and disease. Recent years have seen a great influx of data in glycobiology research with many papers reporting glycosylation sites, proteins, differential glycosylation in reference to disease, etc. However, transferring such data into existing bioinformatics databases for public use requires manual curation by experts, which is often time-consuming and expensive. Accelerating this process is key in facilitating biomedical research by providing the latest findings in a standardized way, ready for research use.

To facilitate Biocuration, we developed a literature mining tool that will detect glycosylated proteins, sites and glycan descriptions automatically in abstracts of publications from the MEDLINE resource. Extracted protein names are normalized to their NCBI gene ID and UniProtKB accession and the data is processed through manual and automatic quality control (QC) checks. The QC process assesses the validity of the protein accessions as well as the reported sites against fasta sequence from the UniProtKB database. The data that passes all QC checks is integrated in the GlyGen (https://www.glygen.org/) database and is publicly available.

Text mining workflow and pipeline



Overview of the process shows how new abstracts and existing ones which have already been collected from UniProtKB, UniCarbKB and GlyTouCan are being mined for annotations.



A multi-step process that involves iterations of refinement, with each step adding to the accuracy of the search parameters and algorithms

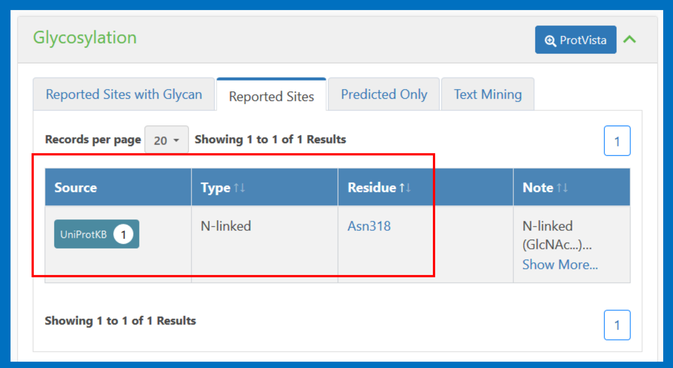
Term	PMID
Bisecting GlcNAc	1374031
Bisialo-biantennary complex type	2386787
core- fucosylated biantennary complex-type oligosaccharide	10731668

Extracted Glycan terms

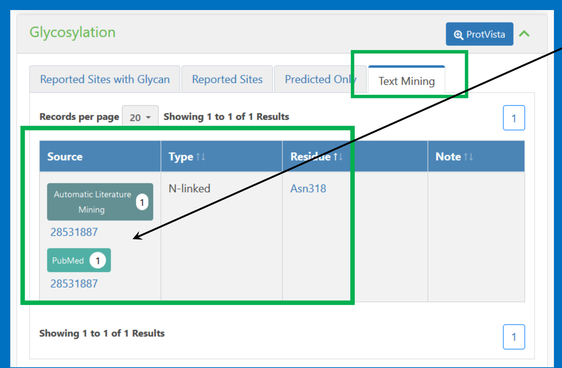
Using chemical and glycobiological descriptions of glycans, a glycan detector was developed – allowed the extraction of motif names, glycan types, sequences etc. This was used to build a glycan dictionary.

Text Mined Results in GlyGen.org

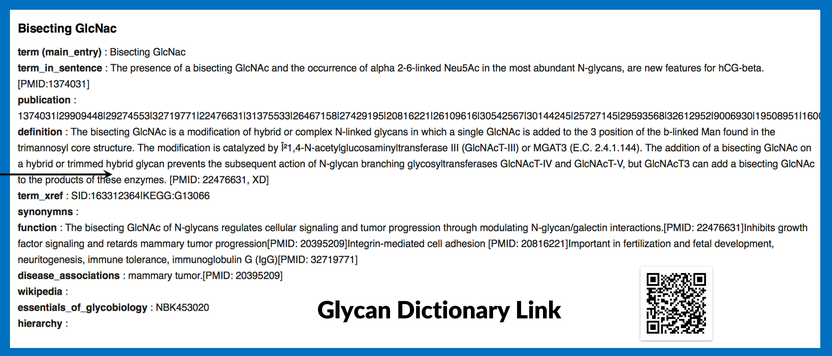
Details For Glycoprotein **Q9ULZ9-1**



Information from UniProt



Text mined information



Bisecting GlcNac

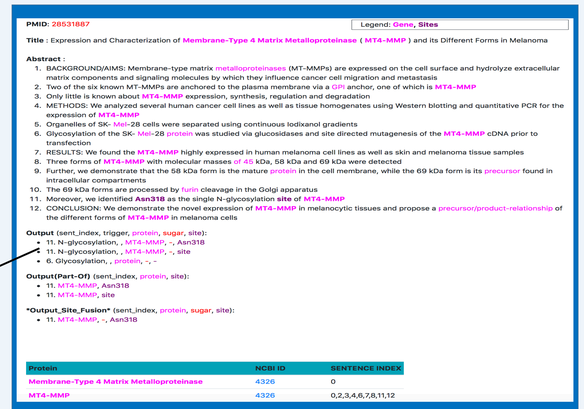
term (main_entry) : Bisecting GlcNac
term_in_sentence : The presence of a bisecting GlcNAc and the occurrence of alpha 2-6-linked Neu5Ac in the most abundant N-glycans, are new features for hCG-beta. [PMID:1374031]
publication : 1374031|2990944|8|2927455|3|2719771|2247663|1|3175533|2646715|8|2742919|5|2081622|1|2610961|6|3054256|7|3014424|5|2572714|5|2953568|3|2612952|9|0069301|1950895|1|160
definition : The bisecting GlcNAc is a modification of hybrid or complex N-linked glycans in which a single GlcNAc is added to the 3 position of the b-linked Man found in the trimannosyl core structure. The modification is catalyzed by β1,4-N-acetylglucosaminyltransferase III (GlcNAcT-III) or MGAT3 (E.C. 2.4.1.144). The addition of a bisecting GlcNAc on a hybrid or trimmed hybrid glycan prevents the subsequent action of N-glycan branching glycosyltransferases GlcNAcT-IV and GlcNAcT-V, but GlcNAcT3 can add a bisecting GlcNAc to the products of these enzymes. [PMID: 22476631, XD]
term_xref : SID:163312364IKEGG:G13066
synonyms :
function : The bisecting GlcNAc of N-glycans regulates cellular signaling and tumor progression through modulating N-glycan/galectin interactions.[PMID: 22476631]Inhibits growth factor signaling and retards mammary tumor progression[PMID: 20395209]Integrin-mediated cell adhesion [PMID: 20816221]Important in fertilization and fetal development, neuritogenesis, immune tolerance, immunoglobulin G (IgG)[PMID: 32719771]
disease_associations : mammary tumor.[PMID: 20395209]
wikipedia :
essentials_of_glycobiology : NBK453020
hierarchy :

Glycan Dictionary Link

Glycan Dictionary

The Human Glycosylation Sites Automatic Literature Mining (https://data.glygen.org/GLY_000481) contains 133 proteins, 241 sites from 160 abstracts.

Acknowledgement -We gratefully acknowledge funding of the GlyGen project by the United States National Institutes of Health (NIH) Common Fund Glycoscience Program (grant U01GM125267-01) through the National Institute of General Medical Sciences (NIGMS).



biotm.cis.udel.edu/glyco/pmid_large/28531887
Extracted text from Abstracts using the Text mining pipeline

Processing Full-Length Articles

Pipeline: List of PMIDs → PMCID → Open Access set → xml files → parse xml retrieve results/conclusion sections and subsections → run glycosylation extraction

Enrichment of Cell Adhesion Molecules and Hematopoietic Cell Linage Markers in Serum Glycoproteins
Except for the component of complement and coagulation cascades, cell adhesion molecules and hematopoietic cell linage markers (CD molecules) were also enriched in the identified serum glycoproteins (supplementary Document S1). Most of these proteins are blood cell or neural cell surface proteins, and they may shed or secreted from cells, such as CD44 protein. Once shedding from cell surface, the soluble CD44 plays versatile biological functions which are different from cell adhesion (27). The N-glycosite N25 of CD44 locates at its extracellular part. One glycopeptide containing this site were identified using the 739 N-glycan masses. (supplemental Table S13). It is known that elevated levels of soluble CD44 in the serum of patients is a marker of tumor burden and metastasis in several cancers (28), hence, the glycans on soluble CD44 may provide more information regarding to disease status. CD44 is also a marker for T cell and erythrocyte lineage differentiation. It is reasonable to detect those CD markers in serum once they shed from progenitor cells into circulation during cell differentiation. The alteration of N-glycosylation on these CD markers may provide information of aberrant cell differentiation.

Modified N-glycans Carried by Serum N-linked Glycoproteins The proposed method will achieve the highest accuracy by using a complete database of N-glycans as well as *N*-linked deglycopeptides derived from human serum. However, this database is not established to date. More than 140 N-glycans in serum was

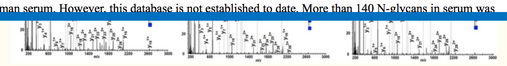


Figure 5

Annotation of nine different glycan structures with varied degree of complexity and sialylation by Mascot on a single glycosylation site (Asn 93) of alpha-1-acid glycoprotein 1 in serum. Shown here are the representative HCD MS2 spectra annotated by Mascot. The nine different glycopeptide variants included the mono-sialylated bi- (**A**), tri- (**B**), tetra-antennary (**C**), and the di-sialylated bi- (**D**), tri- (**E**), tetra-antennary (**F**). Tri- (**G**,**H**) and tetra-sialylated (**I**) glycan structures on the same glycosylation site were also annotated by Mascot.

Development of Protein Name and Detection and Normalization Tool

- Normalization of protein names by in house tool not normalised by PubTator
- Mapping of NCBI ids to UniProt accessions
- Created dictionary from UniProt with special processing for chains and subunits