8-23-2022 10:30 AM

False Discovery Rate Analysis for Glycopeptide Identification

Shun Saito, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Recommended Citation

Saito, Shun, "False Discovery Rate Analysis for Glycopeptide Identification" (2022). *Electronic Thesis and Dissertation Repository*. 8871.
https://ir.lib.uwo.ca/etd/8871

Abstract

Tandem mass spectrometry (MS/MS) is the key technology for glycopeptide identification in high-throughput large-scale glycoproteomics. Estimation of false discovery rates (FDR) is essential for evaluating the quality of the MS/MS-based identification software tools. Although numerous glycopeptide identification tools have been recently proposed, there have been few widely accepted approaches for glycopeptide FDR analysis due to the great structural diversity of glycans. The target-decoy search strategy is currently the most common method for FDR estimation of peptide-spectral matches. In this study, we constructed decoy glycan databases by various methods and compared the FDR from the database search scores produced by each decoy glycan database. Furthermore, we employed a mixture model that facilitates distinguishing between correct and incorrect identifications among the database search score distribution for a better comparison of different decoy glycan database constructions.

Keywords: Tandem mass spectrometry, false discovery rate, target-decoy search strategy

Summary for Lay Audience

Tandem mass spectrometry (MS/MS) is an essential tool to identify chemical substances. Since various glycopeptide identification software have been developed for the past decades, a large quantity of MS/MS data can be identified in a single run of this software. In large-scale glycoproteomics, false discovery rate (FDR) estimation plays a vital role to evaluate the identification results produced by the software because the results may contain incorrect assignments, and manually checking them is not feasible for large datasets. Although extensive research has been carried out on FDR estimation in proteomics, there have been few widely accepted approaches to FDR analysis for glycan because of their structural diversity. Target-decoy search strategy is the standard method to estimate FDR in proteomics, where the sequencing software searches the real target database and incorrect decoy database. In this study, we generated different kinds of decoy glycan databases and compared the effectiveness of the databases for reasonable FDR estimation of glycopeptide identification. To compare the decoy glycan database, we used a mixture model for the differentiation of correct and incorrect glycopeptide assignments.

Acknowlegements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Kaizhong Zhang, without whom I would not have made it through my master's research. His immense knowledge, plentiful experience, and invaluable patience have encouraged me all the time in my academic research life. I have been very fortunate to work with a supervisor like him.

I am also deeply grateful to my thesis examiners, Professor Charles Ling, Professor Boyu Wang, and Professor Xingfu Zou for offering deep insight into my research.

Furthermore, I would like to thank Dr. Baozhen Shan, Dr. Weiping Sun, and Mr. Xiyue Zhang from Bioinformatics Solutions Inc. for providing support for the use of GlycanFinder. I also wish to acknowledge the help provided by the staff in the Department of Computer Science.

Last but not the least, my appreciation also goes to my family and friends for their encouragement during these difficult times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Glycoproteomics

Protein glycosylation is one of the most common and important post-translational modification (PTM). Glycosylation is the covalent attachment of glycans to proteins and occurs not only in animals and plants but also in all other domains of life, such as bacteria and archaea. Protein glycosylation is mediated by glycosyltransferases, the enzymes that catalyze the glycosidic linkages. Interactions between glycosyltransferases, carbohydrate transporters, and glycosidases regulate glycan structures and their functions. Previous research reported that more than 50% of all peptides are modified by glycosylation [1]. It has been estimated that at least 50% and as high as 70% of human proteins are glycosylated [2] and the modification plays a critical role in various biological processes such as protein folding, cell signaling, cellular development, host-microorganism interactions, and immunity. Previous research has found possible relationships between glycosylation and several diseases including Alzheimer's disease and cancers [3] [4].

Glycoproteomics is an actively developing area of research that identifies and characterizes glycosylation at a proteome scale. Glycosylation has not yet been fully understood because, unlike other simple PTMs such as phosphorylation and acetylation, the great structural diversity of glycans and the heterogeneity of glycosylation sites make the glycoproteomics analysis significantly more challenging. A single protein can have hundreds of possible glycan attachments and a specific site of N-linked and O-linked glycosylation can be carried by numerous different glycans [5].

1.1.1 N-linked/O-linked glycans

Various types of glycosylations have been observed that have different target residues for modification and glycosylation sites within proteins. Among them, there are two main types of glycans: N-linked glycans and O-linked glycans. N-linked glycans are attached to asparagine (Asn) side chains in a part of consensus amino acid sequence motif of Asn-X-serine (Ser)/ threonine (Thr), where X can be any amino acid apart from proline (Pro). N-linked glycans commonly have the core glycan structure $GlcNAc_2Man_3$, which is composed of two N-acetylglucosamine (GlcNAc) residues linked to three mannose (Man) residues. Glycans are classified into three groups: high-mannose, complex, and hybrid. High-mannose N-linked glycans have the core structure to which many mannose residues are attached. Complex N-linked glycans consist of the core structure with any type of monosaccharides. Hybrid N-linked glycans have mannose residues on one side of the core structure and complex residues on the other side of the core structure.

O-linked glycans are typically attached to serine (Ser) or threonine (Thr) residues. The most commonly presented monosaccharide is N-acetylgalactosamine (GalNAc) in O-linked glycans, but there is no known consensus amino acid sequence for O-linked glycans as opposed to N-linked glycans. It has been estimated that there are up to 3000 N-linked and O-linked glycans in humans [6]. Aside from GlcNAc, Man and GalNAc, there are several common monosaccharides in N-linked and O-linked glycans such as galactose (Gal), Glucose (Glc), Fucose (Fuc), Xylose (Xyl), N-acetylneuraminic acid (Neu5Ac), and N-glycolylneuraminic acid (Neu5Gc).

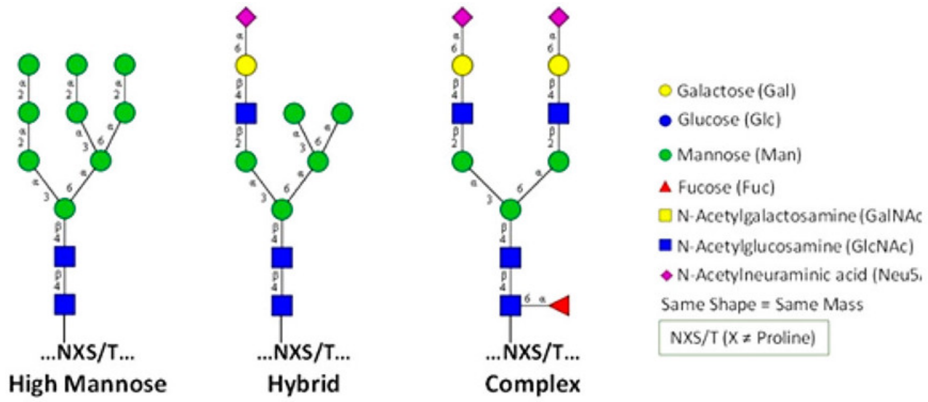


Figure 1.1: Three types of N-glycan [2]

1.1.2 Sequence format for glycans

There are a number of sequence formats to linearly or graphically represent glycan structures. Complex Carbohydrate Structure Database (CCSD, also called CarbBank) [7] uses two-dimensional graphical representations to represent glycan sequences. KCF [8] employs the connection table approach as used by the Kyoto Encyclopedia of Genes and Genomes (KEGG) specifying NODE section for monosaccharides and EDGE section for glycosidic linkages. Glycan data exchange format (Glyde) [9] and CabosML [10] are XML (Extensible Markup Language)-based formats to represent glycomics data regardless of different data acquisition and processing systems. GlycoMinds LinearCode [11], LINUCS applied in GLYCOSCIENCES.de. [12], and Bacterial carbohydrate structure database (BCSDB) [13] represent carbohydrate sequences uniquely with a condensed description of the monosaccharides, connections, and modifications in a linear fashion.

These glycan sequence formats have different ability to encode complex structural features in glycans and none of the current glycan encoding systems can perfectly handle the complex carbohydrate structural data (Table 1.1, where + represents special structual fearures can be encoded, - represents the fearures cannot be encoded, and O means the fearures can be partially encoded). To overcome those limitations of the existing carbohydrate sequence formats, Herget *et al.* proposed GlycoCT format [14].

Sequence format	Multiple connections	Cyclization	Repeating units	Terminal unit underdetermined	Non-stoichiometric modification	Alternative residues
CCSD	–	+	+	o	o	o
LINUCS	–	+	+	–	–	–
BCSDB	–	–	o	–	–	+
LinearCode	–	+	+	+	–	+
KCF	+	+	+	–	–	–
CabosML	–	+	+	–	–	–
Glyde	–	+	+	–	–	–

Table 1.1: Glycan format comparison [14]

GlycoCT

GlycoCT is a connection table-based glycan representation format. GlycoCT employs a graph notation with the residue list containing all the monosaccharides in a glycan and the connectivity list describes all the unique linkages between each monosaccharide (Figure 1.2).

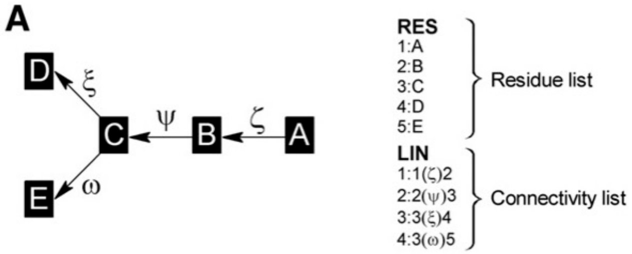


Figure 1.2: GlycoCT general concept [14]

In the residue list, GlycoCT uses five attributes to represent glycan features: anomeric carbon configuration, three-letter stem type code with configuration, the number of carbons in the chain, ring forming positions, and modifier information (Figure 1.3). GlycoCT does not use trivial names like fucose for consistency, but, if needed, these names can be created using other software tools such as GlycanBuilder [15].

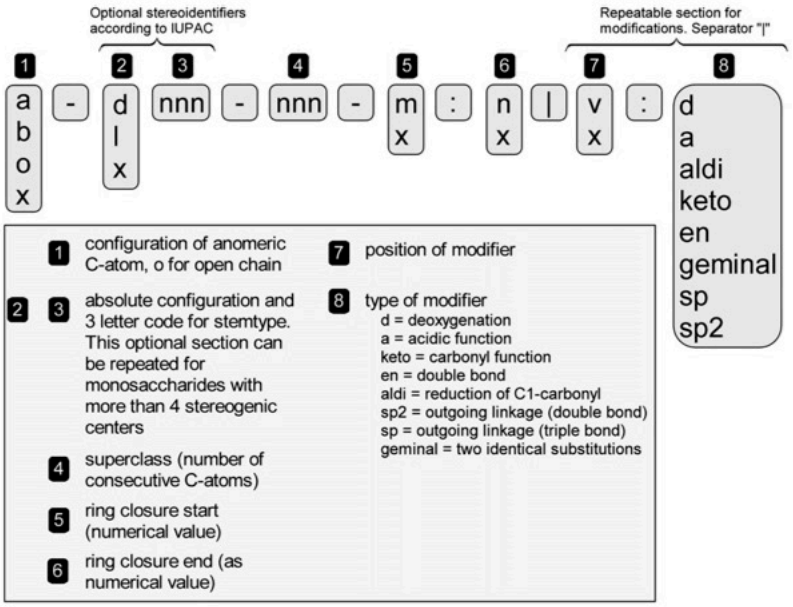

Optional stereoidentifiers according to IUPAC
Repeatable section for modifications. Separator "|"
1
2
3
4
5
6
7
8
a b o x
-
d l x
nnn
-
nnn
-
m x
:
n x
|
v x
:
d a aldi keto en geminal sp sp2
1 configuration of anomeric C-atom, o for open chain
2 3 absolute configuration and 3 letter code for stemtype. This optional section can be repeated for monosaccharides with more than 4 stereogenic centers
4 superclass (number of consecutive C-atoms)
5 ring closure start (numerical value)
6 ring closure end (as numerical value)
7 position of modifier
8 type of modifier
d = deoxygenation
a = acidic function
keto = carbonyl function
en = double bond
aldi = reduction of C1-carbonyl
sp2 = outgoing linkage (double bond)
sp = outgoing linkage (triple bond)
geminal = two identical substitutions

Figure 1.3: GlycoCT residue section [14]

The connectivity list is made up of all the linkages of the monosaccharides in the residue list with the information of residue numbers corresponding to those in the residue list and modification patterns of the linkage by chemical bond formation.

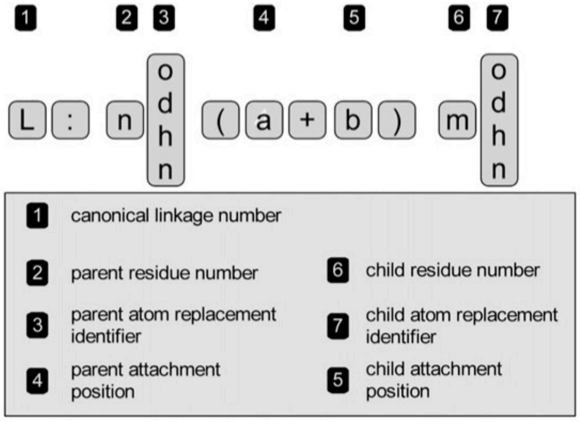

1
2
3
4
5
6
7
L
:
n
o d h n
(
a
+
b
)
m
o d h n
1 canonical linkage number
2 parent residue number
3 parent atom replacement identifier
4 parent attachment position
6 child residue number
7 child atom replacement identifier
5 child attachment position

Figure 1.4: GlycoCT linkage section [14]

1.2 Mass spectrometry

1.2.1 Mass spectrometry

Mass spectrometry (MS) is a widely used method for analyzing complex protein samples. Proteomics based on mass spectrometry is an essential technology for interpreting gene-encoded information. A mass spectrometer basically consists of the ion source, the mass analyzer, and the detector. A mass spectrometer first produces ions from the sample (liquid or gas) in the ion source. Then, it separates ions according to their mass-to-charge ratio, m/z (m means the relative mass of the ions in Daltons (Da), and z means the number of charges, which is counted in accordance with the charge of one electron in absolute value), and fragments the selected ions in the mass analyzer. After that, the mass spectrometer detects the ions, measures the number of ions at each m/z value, which is called abundance, and converts them into electrical signals in the detector. Finally, it processes and records the signals, and transmits them to a computer. The output of a mass spectrometer is a spectrum represented as a set of m/z and intensity pairs. The mass spectrum is mostly shown as a bar graph. The x-axis indicates m/z and the y-axis indicates the intensity, or relative abundance, which is the figure proportional to the number of ions detected in per cent of the most abundant one.

1.2.2 Tandem Mass Spectrometry

The mass analyzer is the fundamental part of the mass spectrometer. Tandem mass spectrometry (MS/MS) is a technique using several analyzers to increase the key parameters in proteomics: the sensitivity, resolution, accuracy of mass measurement, and the ability to generate spectra with a lot of information from peptide fragments. In bottom-up proteomics, proteins are first broken into short peptides by proteases (e.g., trypsin), because the whole protein mass spectrometry is less sensitive than that of the peptide level. Then, the first spectrometer (MS1) selects a precursor ion, which represents a peptide, and fragments it through collision. The second spectrometer (MS2) records and analyzes m/z of each fragmented ion, which is called a product ion. Analyzing the product ions provides information about the peptide sequences

(Figure 1.5).

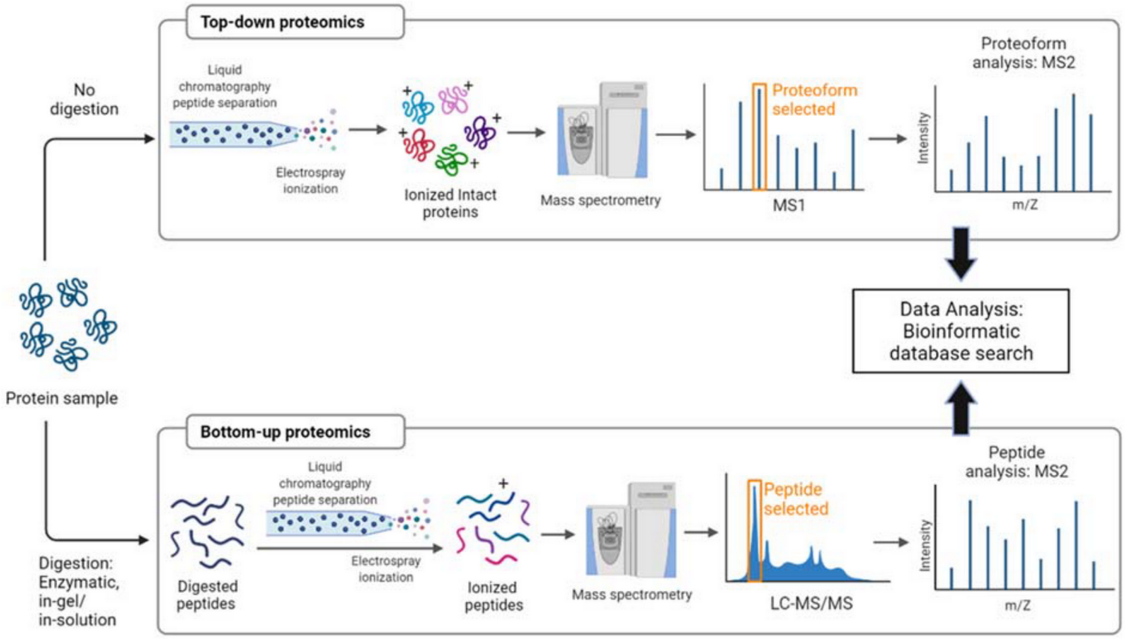


Figure 1.5: Proteomics workflow [16]

1.2.3 Mass spectrometry-based glycoproteomics

Mass spectrometry (MS)-based proteomics has been a standard method for the identification and quantification of glycoproteins thanks to the developments in MS instrumentation, fragmentation strategies, and high-throughput workflows over the past decades. Glycopeptide characterization is often difficult because glycopeptides are low in abundance and commonly used fragmentation methods preferably dissociate glycan with fewer peptide fragments. For these challenges, the traditional approach to identifying glycopeptides is a separation of glycans from peptides by deglycosylation. For N-linked glycopeptides, the enzyme PNGase F is commonly used to remove glycans from glycopeptides [17]. The separated peptides and glycans subsequently are analyzed to identify and characterize peptide sequences, glycosylation sites, and glycan structures. This approach, however, has a drawback in that a glycosylation site and the corresponding glycan structural information cannot be directly obtained. The alternative approach to the deglycosylation strategy is the intact glycopeptide strategy. This strategy leaves glycan attachments to peptides intact, and thus makes it possible to obtain the information for peptide sequences, glycosylation sites, and glycan structure concurrently [18].

1.2.4 Dissociation methods

It is necessary for intact glycopeptide analysis to employ high sensitivity and high throughput tools that can provide fragments from peptide backbones and attached glycans. Several MS peptide fragmentation methods have been developed, which can be classified into two groups by their energy deposition: vibrational methods and electronic methods. The vibrational methods include low-energy collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), and infrared multiphoton dissociation (IRMPD), while the electronic methods include electron capture dissociation (ECD), electron transfer dissociation (ETD) and ultraviolet photodissociation (UVPD).

Using these methods, peptides and glycans are dissociated into many smaller fragment ions. Peptide fragment ions from the C-terminus are labeled as x_1, y_1, and z_1 to x_n, y_n, and z_n, where n is the number of amino acids in the peptide. The other types of peptide fragment ions from the N-terminus are called a_1, b_1, and c_1 to a_n, b_n, and c_n (Figure 1.6). In the same manner, glycan fragment ions from the reducing end are labeled as X_1, Y_1, and Z_1 to X_n, Y_n, and Z_n, from the non-reducing end are A_1, B_1, and C_1 to A_n, B_n, and C_n (Figure 1.7). Different dissociation methods produce different types of fragment ions.

CID [19] is the most employed method to elucidate peptides and glycans. HCD [20] is a higher energy version of CID specific to modern orbitrap mass spectrometers. CID and HCD mostly cleave glycosidic linkages. The single-bond cleavages of precursor ions and low dissociation energy by CID produce abundant B-ions and Y-ions for glycan fragments, and a few b-ions and y-ions for peptide backbone fragments, which are useful information to determine glycan composition, but uninformative for glycosylation sites and peptide sequences (Figure 1.8). In addition to the ions fragmented by CID, HCD can generate several diagnostic oxonium ions, which can be used to distinguish glycan structures. The common oxonium ions are HexNAc internal fragment (m/z = 138.05), Hex (m/z = 163.06), HexNAc (m/z = 204.09), Neu5Ac-H_2O (m/z = 274.09), Neu5Ac (m/z = 292.10), Hex+HexNAc (m/z = 366.14) [21] (Table 1.2). HCD can produce *b*- and *y*- peptide fragment ions, but is likely to lose their glycan modifications. CID and HCD can also produce *A*-ions and *X*-ions resulting from cross-ring fragmentation, which can be informative for glycan structure identification, by modulating the collision energy.

ECD [22] and ETD [23] generate mostly *c*- and *z*- peptide backbone fragments that keep glycan moieties intact, which are used to identify peptide sequences and glycosylation sites. Due to the drawbacks of ETD such as incomplete fragmentation of precursor ions, ETD is often combined with CID/HCD for glycopeptide characterization.

Table 1.2: Monosaccharide mass

Monosaccharide	Abbreviation	Formula	Monoisotopic mass
Galactose	Gal	$C_6H_{12}O_6$	162.0528
Glucose	Glc	$C_6H_{12}O_6$	162.0528
Mannose	Man	$C_6H_{12}O_6$	162.0528
N-Acetylgalactosamine	GalNAc	$C_8H_{15}NO_6$	203.0794
N-Acetylglucosamine	GlcNAc	$C_8H_{15}NO_6$	203.0794
Fucose	Fuc	$C_6H_{12}O_5$	146.0579
Xylose	Xyl	$C_5H_{10}O_5$	132.0423
N-Acetylneuraminic acid	Neu5Ac	$C_{11}H_{19}NO_9$	291.0954
N-Glycolylneuraminic acid	Neu5Gc	$C_{11}H_{19}NO_{10}$	307.0903

1.2.5 Hybrid fragmentation methods

The recent approaches that combine these multiple dissociation techniques are called hybrid fragmentation methods. Hybrid fragmentation methods can complement the limitation of each method and produce more ions and different ion types. In CID/HCD, different fragmentation energy generates different fragmentation types. By stepped collision energy, or SCE, which applies different collisional energy for the same ion groups, more diversified fragmentation ions can be produced [27]. Low energy fragmentation generates *B*-ions, intermediate energy

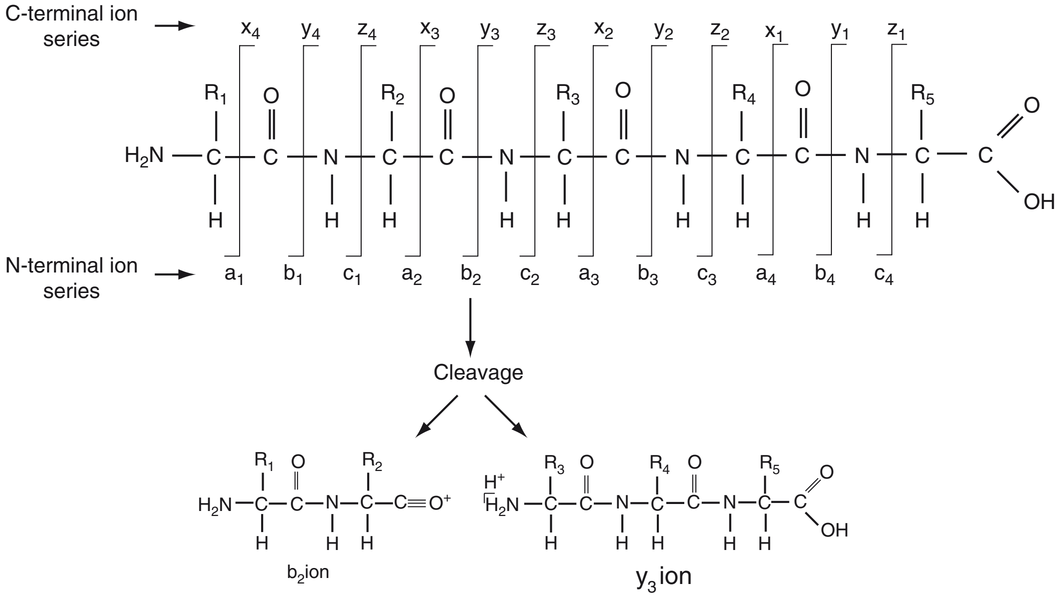

C-terminal ion series
x_4 y_4 z_4 x_3 y_3 z_3 x_2 y_2 z_2 x_1 y_1 z_1
N-terminal ion series
a_1 b_1 c_1 a_2 b_2 c_2 a_3 b_3 c_3 a_4 b_4 c_4
Cleavage
$b_2$ion
$y_3$ion

Figure 1.6: Peptide fragmentation [24]

fragmentation generates Y-ions, and further rounds of high energy fragmentation on Y_1 ions for N-glycopeptides or Y_0 ions for O-glycopeptides generate peptide backbone fragments. Previously, multiple rounds of fragmentation on the same or different collisional energy were performed sequentially, which tends to take a longer time, but the recent development of the instruments has made it possible in one scan.

Product-ion triggered fragmentation combines different fragmentation methods on the same precursor ions [28]. HCD can accurately detect low m/z fragment ions, and thus it is used to search for the presence of glycans by diagnostic oxonium ions. Since the most abundant fragment ion in N-glycopeptides is HexNAc, HexNAc oxonium ion (m/z = 204.09), its internal fragment ions (m/z = 138.05 and 168.07), and the combination of ions at these m/z values are typically used to detect the diagnostic oxonium ions. If a diagnostic oxonium ion is detected, an additional round of ETD fragmentation is carried out, which is called HCD-product dependent-ETD or HCD-pd-ETD [29].

Electron transfer/higher-energy collisional dissociation (EThcD) [30], a combination of the collisional dissociation method and the electron dissociation method, produces abundant structural information on glycopeptides with both glycan fragment ions and peptide backbone

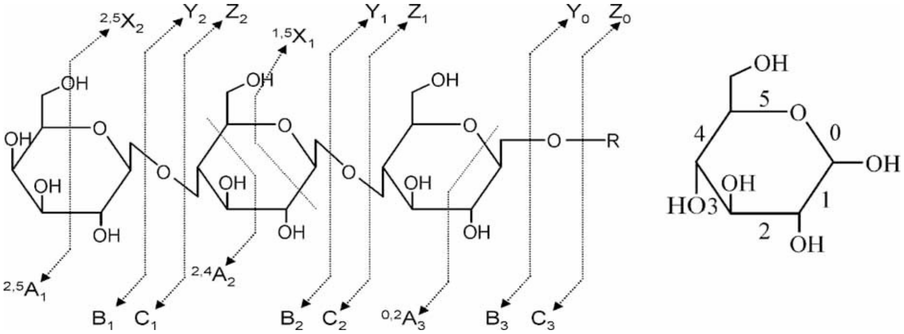


Figure 1.7: Glycan fragmentation [25]

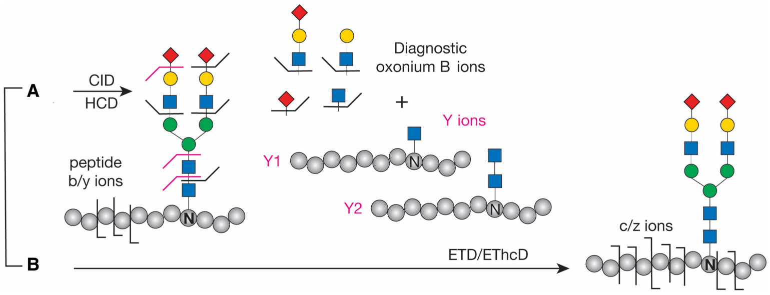


Figure 1.8: Glycopeptide fragmentation by diffrerent dissociation methods [26]

fragment ions in one single spectrum. Previous research shows EThcD outperforms ETD and HCD for larger glycopeptides [31].

1.2.6 Peptide identification strategy

There are three common strategies for peptide identification: database searching, spectral library searching, and *de novo* sequencing. Database searching [32] is the most widely used strategy for peptide identification and characterization in bottom-up proteomics. The strategy is to match experimental MS/MS data with a theoretically possible sequence in reference proteome databases including UniProtKB [33] and NCBI RefSeq [34]. If proteases have a known digestion pattern and peptides have a known fragmentation pattern, a list of plausible peptides and corresponding fragments is produced. By computationally comparing the experimental mass spectrum with the theoretical fragment masses (Table 1.3), peptides are scored and ranked depending on the degree of matching between candidate peptides and the experimental data, and the best-scoring peptide is reported. The most common search engines for database searching are SEQUEST [35] and MASCOT [36].

Database searching has drawbacks in that the strategy heavily depends on the quality and availability of reference databases. When an organism of interest has not been sequenced, or when there are no accurate reference databases because of splice variants, single amino acid variations and PTMs, database searching does not work well. Database searching also has some disadvantages such as false positive identifications caused by noisy spectra and scoring imbalances between low-quality long peptides and high-quality short peptides. One of the alternative strategies to database searching is a direct spectrum-to-spectrum comparison between experimental MS/MS spectra and reference.

MS/MS spectra in a spectral library is referred to as spectral library searching [37]. Although spectral library searching has lower processing times and potentially higher identification rates compared to database searching, it also relies on available accurate reference databases for spectrum data. To compare theoretical and experimental data, a sufficient amount of precisely annotated MS/MS spectra is needed in spectral libraries.

When there is no appropriate database and to overcome those disadvantages of the database-dependent approach, *de novo* sequencing [38] approach is the only way for peptide identification. *De novo* sequencing can reconstruct the original amino acid sequences from an MS/MS spectrum and make it possible to identify previously unknown peptide sequences, peptide homologues, and modifications. Also, the results of *de novo* sequencing can be used to validate the results of database searching, because both results are very similar [39]. There are various widely used *de novo* sequencing software available such as PEAKS [40], Novor [41], and PepNovo [42].

Table 1.3: Amino acid mass

Residue	3-letter code	1-letter code	Formula	Monoisotopic mass
Alanine	Ala	A	C_3N_5NO	71.03712
Arginine	Arg	R	$C_6H_{12}N_4O$	156.10112
Asparagine	Asn	N	$C_4H_6N_2O_2$	114.04293
Aspartic acid	Asp	D	$C_4H_5NO_3$	115.02695
Cysteine	Cys	C	C_3H_5NOS	103.00919
Glutamine	Gln	Q	$C_5H_8N_2O_2$	128.05858
Glutamic acid	Glu	E	$C_5H_7NO_3$	129.04260
Glycine	Gly	G	C_2H_3NO	57.02147
Histidine	His	H	$C_6H_7N_3O$	137.05891
Isoleucine	Ile	I	$C_6H_{11}NO$	113.08407
Leucine	Leu	L	$C_6H_{11}NO$	113.08407
Lysine	Lys	K	$C_6H_{12}N_2O$	128.09496
Methionine	Met	M	C_5H_9OS	131.04049
Phenylalanine	Phe	F	C_9H_9NO	147.06842
Proline	Pro	P	C_5H_7NO	97.05277
Serine	Ser	S	$C_3H_5NO_2$	87.03203
Threonine	Thr	T	$C_4H_7NO_2$	101.04768
Tryptophan	Trp	W	$C_{11}H_{10}N_2O$	186.07932
Tyrosine	Tyr	Y	$C_9H_9NO_2$	163.06333
Valine	Val	V	C_5H_9NO	99.06842

1.3 False discovery rates and target-decoy approach

In proteomics, a database search algorithm is used to obtain confidence metrics such as p-value and e-value after examining spectra against peptides in a database. The algorithm performs verification for a single peptide identification using these metrics. These metrics represent the goodness of fit of an observed spectrum and the corresponding peptide candidate, but most search engines assign all experimental MS/MS spectra to peptides in a database if they are within specified mass tolerance. It has previously been observed that only 10-50% of spectrum assignments are correct in MS/MS experiments [43]. This is because not all peptides are included in the reference database and incorrect peptide candidates sometimes can outscore correct sequences. For small datasets, it is reasonable to manually examine each PSM to verify identification correctness, but in the current large-scale, high-throughput proteomics, this strategy is not viable. For a large group of identification, rather than examining the correctness of each assignment, the proportion of incorrect identifications is estimated for assignment verification, which is called false discovery rate (FDR) estimation.

1.3.1 Target-Decoy Search Strategy

Large-scale proteomics requires a method to estimate the proportion of incorrect peptide assignments among correct assignments. Target-decoy search strategy [44] is simple to implement and a standard strategy to estimate FDR in large-scale proteomics. To estimate FDR in the target-decoy search strategy, decoy peptide sequences that do not exist in nature are created. The target-decoy search strategy assumes that the original peptide database (target database) and the decoy database do not overlap so that decoy hits are incorrect assignments. Decoy sequences, therefore, should be constructed to avoid the common peptide sequences between the target and the decoy database and to preserve the general composition of the sequences in the target database.

The other assumption of this strategy is that false positive identifications are equally likely to come from the target database and the decoy database. Incorrect decoy peptides should be similar to incorrect but unknown peptides derived from target peptides regarding peptide length, amino acid composition, peptide mass, and output scores from the search engine. Experimental MS/MS spectra are then searched against the target and decoy database. Since peptide sequences in the decoy database cannot exist in the sample, any PSMs to the decoy sequences are incorrect identifications and one can estimate the relative proportion of target and decoy sequences.

1.3.2 Decoy sequence construction

Several methods for constructing decoy sequences have been developed. Each method has its advantages and disadvantages, and there is no single best way to create a decoy database. Reversing the amino acid sequences in the target database is one of the simplest and most common ways to create decoy sequences. This method has advantages in that the general features of the target sequence such as peptide length and amino acid composition are preserved. Also, this reversal method is so simple to implement that other researchers can create the same decoy sequences. On the other hand, this method has disadvantages in that decoy sequences by reversal are not random transformations and it is difficult to create decoy peptides corresponding to the target sequences with palindromic or low complexity.

Shuffling sequences is another method for decoy sequence construction. Since this method randomly rearranges the target peptide sequence, it is easy to deploy and preserves peptide length and amino acid composition of the target sequences like the reversal method.

One can also completely randomize the sequences to generate decoy sequences. Randomized sequences should preferably have the same distribution of peptide length and amino acid composition as those in a target database. For that purpose, the random method first creates a frequency matrix of amino acids and a histogram of peptide length in the target database, and then randomly chooses amino acids according to the frequency matrix, and adds these amino acids until a specified length. The shuffle method and the random method have a drawback in that they do not preserve redundancies and homologies between peptide sequences, and thus there can be much more decoy sequences than sequences in the target database. For FDR estimation, this imbalance and observed decoy bias should be considered.

There are two main types to carry out target-decoy search (Figure 1.9). Separate database search is performed by searching the target database and the decoy database separately. In the separate search, two identifications are reported for each spectrum: the target identification from the best score in the target database and the decoy identification from the best score in the decoy database. When searching the target and the decoy database separately, there is no competition between target and decoy sequences for the top-ranked score in a single search. Some researchers argue that decoy sequences that partially match high-quality MS/MS spectra are likely to get higher scores than other top-ranked matches in the separate database search, and thus higher score threshold should be used [44]. Also, separate database search has difficulty in estimating correct identifications with the low scores because of incorrect identifications with high scores.

Concatenated database search is performed by combining a database of the target and decoy sequences. In the concatenated search, only one match with the best score from either target or decoy sequences is reported for each spectrum based on the idea that when a given PSM is correct, the target sequence is expected to produce a higher score than the decoy sequence. On the other hand, when a PSM is not correct, there is an equal probability of matching a target sequence and a decoy sequence.

1.3.3 FDR estimation

In the context of the target-decoy search strategy, true positive (TP) means the number of correct assignments above a given score threshold, whereas false positive (FP) means the number of incorrect assignments above a given score threshold. True negative (TN) represents the number of incorrect assignments below a given score threshold, whereas false negative (FN) represents the number of correct assignments below a given score threshold.

Sensitivity refers to the fraction of all correct assignments above a given score threshold, and using the above notations, it can be written as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.1}$$

On the other hand, specificity represents the fraction of all incorrect assignments above a given score threshold written as

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1.2}$$

Precision is the fraction of correct assignments above a given score threshold that is calculated by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.3}$$

FDR is the fraction of incorrect assignments above a given score threshold, which estimates the ratio of the incorrect PSMs among the accepted PSMs in the target-decoy search strategy, written as

$$\begin{aligned} \text{FDR} &= 1 - \text{Precision} \qquad (1.4) \\ &= \frac{\text{FP}}{\text{TP} + \text{FP}} \qquad (1.5) \end{aligned}$$

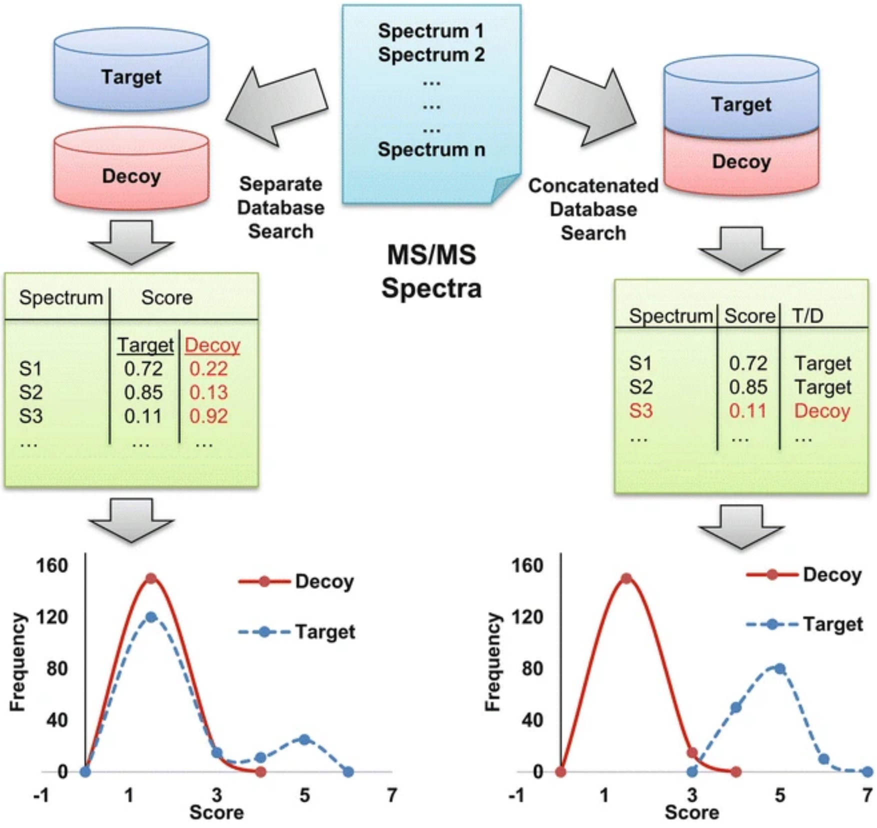
Separate/simple FDR simply computes the ratio of the number of decoy PSMs above the threshold and the number of target PSMs above the threshold for a given score threshold:

$$\text{FDR} = \frac{\text{Number of decoy PSMs above the threshold } (D)}{\text{Number of target PSMs above the threshold } (T)} \tag{1.6}$$

where decoy PSMs are reported by searching a decoy peptide database and target PSMs are reported by searching the original peptide database.

Concatenated database search assumes that there is the same number of false identifications in target PSMs above a given threshold as the number of decoys above the threshold, and therefore, the number of false positives is doubling the number of decoy PSMs above the threshold. The true positive PSMs tend to match the target sequences, while the false positive PSMs are equally distributed among target and decoy sequences. The number of decoy PSMs represents half the number of false positive PSMs.

$$\text{FDR} = \frac{2 \times D}{T + D} \tag{1.7}$$



Spectrum	Score	
	Target	Decoy
S1	0.72	0.22
S2	0.85	0.13
S3	0.11	0.92
...	...	...

Spectrum	Score	T/D
S1	0.72	Target
S2	0.85	Target
S3	0.11	Decoy
...	...	...

Figure 1.9: Schema of separate database search and concatenated database search [45]

Käll *et al.* [46] proposed a more sophisticated calculation method than simple separate FDR by incorporating the percentage of incorrect target PSMs (PIT). Separate FDR does not consider incorrect target PSMs. The target-decoy search strategy assumes not all target PSMs are correct while all decoy PSMs contribute to incorrect matches, and thus, the set of target PSMs consists of a mixture of correct and incorrect target PSMs. This bimodal target score distribution containing correct and incorrect target hits causes an overestimation of FDR. To consider this bias, these incorrect target PSMs need to be factored into the calculation of FDR. To estimate the PIT, which is also commonly known as π, from the observed score distributions, PIT is calculated by the ratio of the number of false discoveries to the total number of PSMs. For example, supposing 1000 target PSMs, which contain 200 correct PSMs and 800 incorrect PSMs, PIT equals 0.8. Then, the FDR of this method is the ratio of the number of decoy PSMs to the number of target PSM multiplied by PIT, 0.8 in this example. This method needs estimation of PIT based on the experimental score distributions. Factoring into this correction weight allows FDR estimation more accurate.

$$\text{FDR} = \text{PIT} \times \frac{D}{T} \tag{1.8}$$

Extensive research has been carried out to examine the methods for FDR estimation for peptide identification, but there have been no common methods to calculate FDR for glycopeptide identification because of the great diversity of glycan structures. In this thesis, we carry out various approaches for decoy glycan database construction based on target-decoy search strategy for validation of peptide identification. However, the target-decoy strategy for peptide identification and that for glycopeptide identification is different in that peptides consists of a linear sequence of amino acids whereas glycans are composed of monosaccharide in a tree structure. One of the strategies employed in previous research is generating theoretical target glycopeptide spectra and adding a random mass to Y-ions of the target spectra, which yield decoy spectra. We constructed decoy glycan databases by maintaining the tree structure of the glycans and changing the monosaccharides in the tree. In addition, we employ a mixture model that facilitates distinguishing between correct and incorrect identifications among the database search score distribution for a better comparison of different decoy glycan database construction and examines the appropriateness and effectiveness of the simple FDR estimation method.

Chapter 2

Methods

2.1 Datasets

In this work, publicly available RAW MS data of mouse brain glycopeptide samples were retrieved from PRIDE proteomics database [47]. This dataset was analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) using an Orbitrap mass analyzer with HCD fragmentation. The protein database was obtained from UniProtKB/Swiss-Prot [48]. The following parameters of the database search software were used in our experiment. The precursor mass tolerance was ±10 p.p.m., the fragment ion mass tolerance for peptide was ±0.2 *Da*, and the fragment ion mass tolerance for glycan was ± 20 p.p.m. Trypsin was used as the enzyme for protein digestion. PTMs were specified for glycan search. Carbamidomethylation on cysteine residues ($C + 57.02\ Da$) was set as the fixed modification, in which all cysteine residues were modified. Oxidation on methionine residues ($M + 15.99\ Da$) was set as the variable modification, some of which were modified. There were 2794 N-glycans in our target glycan database.

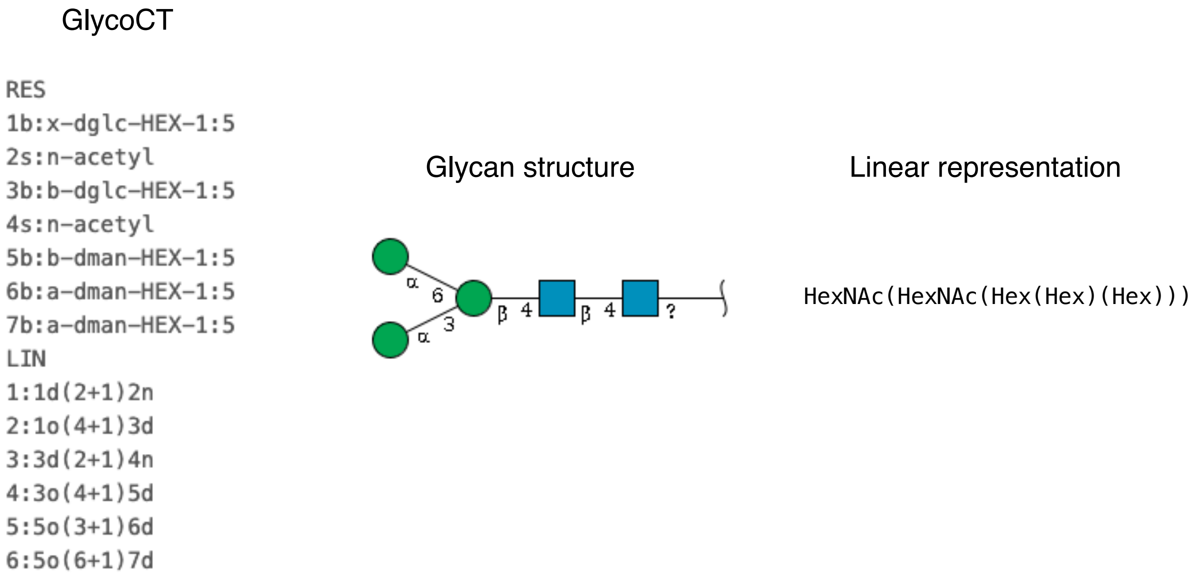
2.2 Notations

There are various kinds of monosaccharides, but in this study, we considered six frequently observed monosaccharide residue types: Hex, HexNAc, Fuc, Xyl, NeuAc, and NeuGc (Table 2.1). Since glucose, mannose, and galactose have the same formula and mass, they are classified into Hex in this work. In the same way, N-Acetylglucosamine and N-Acetylgalactosamine are categorized as HexHAc.

Table 2.1: Types of monosaccharides

Monosaccharide	Symbol	Monoisotopic mass	Generic term
Galactose	🔵	162.0528	Hex
Glucose	🟡	162.0528	Hex
Mannose	🟢	162.0528	Hex
N-Acetylgalactosamine	🟨	203.0794	HexNAc
N-Acetylglucosamine	🟦	203.0794	HexNAc
Fucose	🔺	146.0579	Fuc
Xylose	⭐	132.0423	Xyl
N-Acetylneuraminic acid	◆	291.0954	NeuAc
N-Glycolylneuraminic acid	◇	307.0903	NeuGc

There were 2794 glycans in our glycan database in which each glycan is represented by GlycoCT connection table-based format. To construct decoy glycan databases, glycans in GlycoCT format were transformed into glycans represented by the linear notation (Figure 2.1).



```
RES
1b:x-dglc-HEX-1:5
2s:n-acetyl
3b:b-dglc-HEX-1:5
4s:n-acetyl
5b:b-dman-HEX-1:5
6b:a-dman-HEX-1:5
7b:a-dman-HEX-1:5
LIN
1:1d(2+1)2n
2:1o(4+1)3d
3:3d(2+1)4n
4:3o(4+1)5d
5:5o(3+1)6d
6:5o(6+1)7d
```

```
HexNAc(HexNAc(Hex(Hex)(Hex)))
```

Figure 2.1: Glycan tree example

2.3 Decoy glycan construction

To construct a decoy glycan database, we rearranged nodes of a glycan in the target database while keeping the target glycan structure. In addition, we generated decoy glycan sequences so that there were no common sequences between the target glycan database and the decoy glycan database, which conforms to the key assumption of the target-decoy search strategy.

2.3.1 Permutation method

This method permutates all the monosaccharides in a given glycan (Figure 2.2). After carrying out multiple sets of permutations for each glycan, we selected a decoy glycan sequence that had a maximum tree edit distance [49] from the corresponding target glycan sequence. The glycan topology and the composition of monosaccharides in a given target sequence were preserved for the corresponding decoy glycan in this method. We created decoy glycan databases from 10 sets, 30 sets, and 60 sets of permutation.

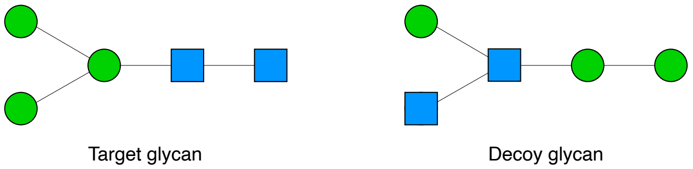


Figure 2.2: Permutation method

2.3.2 Swap method

The swap method replaces nodes between a pair of monosaccharides, which has the effect of randomly adding or subtracting a certain amount of mass to the corresponding target glycan. First, we made three pairs of monosaccharides that have similar residue mass (Table 2.1): Hex and HexNAc, Fuc and Xyl, and NeuAc and NeuGc. Then, a monosaccharide was swapped with the other monosaccharide in the pair. For example, if a Hex node is presented at a given position of the tree structure in a target glycan, the Hex node is changed to a HexNAc node on the same position of the same tree structure in the corresponding decoy glycan (Figure 2.3). In this method, glycan topology is preserved but monosaccharide composition and glycan mass are different between the target glycan and the corresponding decoy glycan. We created swap-based decoy databases where 50%, 75%, and 100% of monosaccharides in a target glycan were swapped to generate the corresponding decoy glycan.

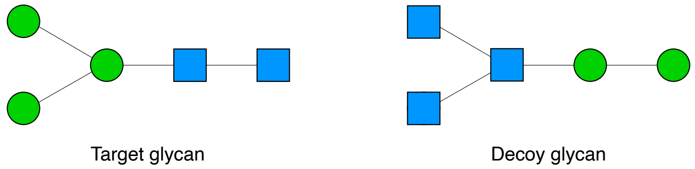


Figure 2.3: Swap method

2.3.3 Random method

For the random method, we first observed the frequency of each monosaccharide in the whole target database: HexNAc, 0.380; Hex, 0.409; Fuc, 0.103; Xyl, 0.001; NeuAc, 0.058; and NeuGc, 0.049. We then randomly chose monosaccharides for each node in the glycan tree in accordance with the monosaccharide frequency in the target database (Figure 2.4). Similar to the swap method, monosaccharide composition and glycan mass in the target glycan sequence were not preserved for the corresponding decoy glycan sequence, although the tree structure for the decoy glycan was maintained in the random method.

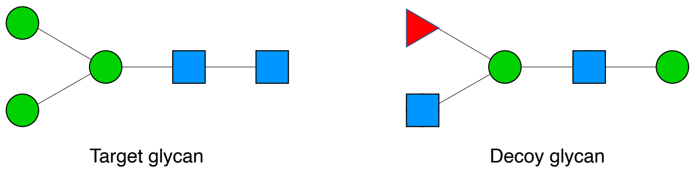


Figure 2.4: Random method

2.4 Software

GlycanFinder is an advanced feature of PEAKS Studio [50], which is a software platform that performs protein identification and quantification, PTM analysis, and peptide *de novo* sequencing, that is designed to identify and quantify glycopeptides from LC-MS/MS spectra data. The software tool provides identification of glycosylation sites and visualization of the N-linked and O-linked glycan structures, annotation of spectra, and glycan *de novo* sequencing for a comprehensive analysis and interpretation of glycopeptides. The software is a preliminary version developed by Bioinformatics Solutions Inc. GlycanFinder has the default built-in glycan database, but we can incorporate our own decoy glycan database into the software and compare the results from different decoy glycan databases. Although the new version of this software has been developed recently, we show the results from the previous version we have been using in this thesis. For the default decoy method, the latest version of GlycanFinder adds random mass to the mass of theoretical target spectra to generate the decoy spectra and search experimental spectra against target and decoy spectra.

2.5 FDR estimation

We calculated FDR by the following simple calculation method:

$$\text{FDR} = \frac{\text{Number of decoy PSMs above the threshold } (D)}{\text{Number of target PSMs above the threshold } (T)} \tag{2.1}$$

From all the glycopeptide-spectrum matches (GPSMs), FDR for each GPSM was calculated using Equation 2.1. In this work, we focused on the analysis of glycan FDR, which was obtained by glycan target matches and glycan decoy matches, and analyzed the target glycan matches and decoy glycan matches below 1% glycan FDR.

2.6 Comparative search engine

We compared the above decoy construction methods using a different search engine for glycopeptide identification: GlycanFinder and pGlyco 2.0 [51]. pGlyco 2.0 is one of the most common software tools for intact glycopeptide identification, which conducts glycan-level, peptide-level, and glycopeptide-level false discovery rate evaluation for glycopeptides. pGlyco 2.0 employs its decoy method based on a mass list of glycopeptides and random addition of mass. GlycanFinder and pGlyco 2.0 used the same parameter setting for the analysis of mouse brain glycopeptides.

2.7 Mixture model

From the assumption of the target-decoy strategy that the target database and the decoy database do not have any overlapped entries, all the decoy hits are incorrect matches. On the other hand, target hits can contain both correct and incorrect matches. Therefore, the distribution of glycan scores for target matches can be considered as a mixture of a distribution of glycan scores for correct matches and a distribution of scores for incorrect matches. The objective of the mixture model approach is to estimate the distribution parameters from the observed data. To better assess the effectiveness of glycopeptide identification software, we use a statistical model for a distinction between correct identifications and incorrect identifications among target matches. We modeled the distribution of glycan score data with a mixture of two component distributions representing correct score distribution for one and incorrect score distribution for the other by the Bayesian approach.

It is unknown from which mixture component each observed data comes. The mixture of multiple distributions is a weighted sum of K components formulated by

$$f(x;\theta_1,...,\theta_k) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \tag{2.2}$$

where θ_k represents parameters of each component in the mixture and the mixing weights π_k

meet the conditions

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.3}$$

and

$$\pi_k \geq 0 \tag{2.4}$$

The distributions can be the same parametric family of distributions such as normal with different distribution parameters or different distributions.

2.7.1 Bayesian approach

In Bayesian statistics, probability reflects a degree of belief in a hypothesis, and the parameter θ is modeled as a random variable unlike the frequentist approach, where it determines the parameter θ that represents the true distribution of data. At an initial stage, probability, in which prior knowledge about parameters from previous experiences is included, is subjective to some degree. Then, the degree of belief is updated while observing data. Using this method, we can infer the parameter θ by producing a probability distribution for θ. Point estimate can be extracted from the distribution. In the Bayesian method, using a probability density function $P(\theta)$, which is called prior distribution of the parameter θ, that shows the degree of belief about parameter θ and a statistical model $P(x|\theta)$, which is called likelihood that represents the belief about x given θ with an observation of data $x_1, x_2, ..., x_n$, the belief is updated, and $f(\theta|x_1, x_2, ..., x_n)$, which is called posterior distribution, is calculated.

The joint probability mass function for θ and x is defined as

$$P(\theta, x) \quad = \quad P(x|\theta)P(\theta) \tag{2.5}$$

Using the conditional probability

$$P(\theta|x) \quad = \quad \frac{P(\theta, x)}{P(x)} \tag{2.6}$$

and the law of total probability

$$P(x) \quad = \quad \sum_{i=1}^{n} P(x_i|\theta)P(\theta) \tag{2.7}$$

Bayes theorem is defined as

$$P(\theta|x) \quad = \quad \frac{P(x_i|\theta)P(\theta)}{\sum_{j=1}^{n} P(x_i|\theta)P(\theta)} \tag{2.8}$$

Since the denominator in equation (2.8) does not depend on θ, and it can be considered a constant, equation (2.8) can be written as

$$P(\theta|x) \quad \propto \quad P(x|\theta)P(\theta) \tag{2.9}$$

where the symbol $\propto$ represents proportionality. Bayesian methods carry out the computation to yield $P(\theta|x)$, which represents the updated belief after observing data.

2.7.2 Mixture model for glycopeptide score distribution

Since we do not have labels of correct assignments and incorrect assignments in the target distribution, this mixture model is unsupervised learning that extracts useful information from unlabeled data for the distinction between correct and incorrect identifications. The score distribution of the glycopeptide assignment was modeled by a one-dimensional mixture model, in which the score data was one-dimensional with a single variable. The distribution consists of two component distributions $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$ for correct assignment score distribution and incorrect assignment score distribution, respectively. This two-component mixture model can be written by

$$f(x;\theta_1,\theta_2) \quad = \quad \pi_1 f_1(x;\theta_1) + \pi_2 f_2(x;\theta_2) \tag{2.10}$$

where $\pi_1 + \pi_2 = 1$. The likelihood and log-likelihood of this mixture model are represented by

$$\begin{aligned} L(\theta_1,\theta_2) &= \prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) & (2.11)\\ &= \prod_{i=1}^{n}\Big(\pi_1 f_1(x_i;\theta_1) + \pi_2 f_2(x_i;\theta_2)\Big) & (2.12)\end{aligned}$$

$$l(\theta_1,\theta_2) \quad = \quad \sum_{i=1}^{n} \log\Big(\pi_1 f_1(x_i;\theta_1) + \pi_2 f_2(x_i;\theta_2)\Big) \tag{2.13}$$

where $\pi_1 + \pi_2 = 1$. To accurately calculate the probability that glycopeptides are correctly or incorrectly assigned, models of glycan score distributions for correct and incorrect assignment are needed. From the empirical observation, the distribution of correct glycopeptide assignment was modeled by normal distribution and the probability of correct glycopeptide identification having database search score S can be calculated by

$$p(S|+) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{S-\mu}{\sigma})^2} \tag{2.14}$$

with mean μ, standard deviation σ, and + that represents correct glycopeptide assignment, On the other hand, lognormal distribution can closely approximate the observed incorrect glycopeptide identifications because the shape of the distribution was asymmetric and had a long right tail. The probability of incorrect glycopeptide identification having a database search score S can be calculated by

$$p(S|-) = \frac{1}{\sigma\sqrt{2\pi}S} e^{-\frac{1}{2}(\frac{\log S-\mu}{\sigma})^2} \tag{2.15}$$

with mean μ, standard deviation σ, and $-$ that represents incorrect glycopeptide assignment.

Using the Bayes theorem, the probability of correct glycopeptide assignment to spectrum i with database search score S is calculated by

$$p(+|S_i) = \frac{\pi_1 f_1(S_i)}{\pi_1 f_1(S_i) + \pi_2 f_2(S_i)} \tag{2.16}$$

where + represents the correct glycopeptide assignment. The log-likelihood of the mixture distribution of glycopeptide search score is described by

$$\begin{aligned} l &= \sum_{i=1}^{n} \log f(S_i) && (2.17) \\ &= \sum_{i=1}^{n} \log\Big(\pi_1 f_1(S_i) + \pi_2 f_2(S_i)\Big) && (2.18) \end{aligned}$$

2.7.3 Markov chain Monte Carlo (MCMC)

To distinguish correct and incorrect identifications, we employed Markov chain Monte Carlo (MCMC) based mixture model. MCMC is a popular method to determine probability density function parameters by repeatedly generating samples using Markov chain to find best-fitting values. The MCMC approach on Bayesian models is a fast and flexible method, especially when there are numerous parameters, that can be applied to a wide range of problems. The MCMC-based method does not require evaluation of the likelihood functions, which may have a large number of integrals. As opposed to maximum likelihood parameter estimation using numerical optimization, the MCMC generates a sample of parameter values from the posterior distribution of the model parameters. After a multitude of iterations is carried out, the posterior distribution is yielded from the sample distribution of the parameters.

The Metropolis algorithm [52] is one of the basic sampling methods for MCMC. The Metropolis algorithm performs random walks with an acceptance/rejection rule to converge to the desired target distribution. The algorithm starts with an arbitrarily chosen starting point of the model parameters, $\theta^0 = (\theta_1^0, ..., \theta_m^0)$ from starting probability density $p_0(\theta)$. For each iteration $t = 1, 2, ...$, a candidate for the next sample value θ^* is generated from proposal distribution at time t, $g_t(\theta^*|\theta^{t-1})$. The proposal distribution can be chosen depending on the current state θ^{t-1}, according to the Markov chain property. Then, the acceptance ratio to decide whether the candidate is accepted or rejected is calculated by

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} \tag{2.19}$$

After sampling the random candidate θ^*, whether the candidate is accepted or rejected is decided.

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases} \tag{2.20}$$

The algorithm calculates the ratio r for all (θ, θ^*), and samples θ from the proposal distribution $g_t(\theta^*|\theta)$ for all θ and t.

The Metropolis-Hastings [53] algorithm is a generalized method of the Metropolis algorithm. For the Metropolis algorithm, the proposal distribution has to be symmetric, which must meet the condition $g_t(\theta_a|\theta_b) = g_t(\theta_b|\theta_a)$, but this condition is not required for the Metropolis-Hastings algorithm. Thus, the acceptance ratio is rewritten as

$$r = \frac{p(\theta^*|y)/g_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/g_t(\theta^{t-1}|\theta^*)} \tag{2.21}$$

Due to the asymmetric jumping rules, the Metropolis-Hastings can perform random walks more efficiently than the basic Metropolis algorithm. This algorithm Converges to the desired target distribution in the same manner as the Metropolis algorithm does.

Using the MCMC algorithm, we differentiate correct and incorrect glycopeptide assignments from a mixture of two component distributions. Three chains were generated so that the algorithm had different starting values and evaluate the convergence of MCMC. The first 1000 samples obtained during the adaptive phase were discarded, in which a Markov chain is not formed. The number of iterations to reach convergence and achieve correctness in the sampling phase was set to 3000. To control larger jumps in the chain and keep sample values to be close to the previous samples, the MCMC algorithm kept every 5 sampled values and discarded other samples. After separating the two distributions, we compare the shape of the decoy distribution, which is incorrect score distribution based on the key assumption of the target-decoy search strategy, and the shape of the target incorrect distribution, which is separated from the whole target assignment score distribution.

2.7.4 Fitting decoy glycan score distribution

We fitted a probability distribution to data from decoy glycan scores by using maximum likelihood estimation. Let data x and the probability of observing data $P(x|\theta)$. Since x is known and the parameter θ is unknown, the value of $P(x|\theta)$ is a function of θ, which is called the likelihood of the data x and is denoted as $L(\theta; x)$. Let $x = (x_1, ..., x_n)$ be a sample independently observed from a distribution, then

$$L(\theta; x) = \prod_{i=1}^{n} P(x_1|\theta) \tag{2.22}$$

and the estimate of the parameter is

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x) \tag{2.23}$$

The lognormal distribution is used to model continuous random variables greater than or equal to zero. Another characteristic of the lognormal distribution is the distribution is skewed to the right. By observing the shape of the distribution of the decoy glycan score, we assumed the lognormal distribution can be better fitted to the data since the lognormal distribution has a heavier right tail and lighter left tail compared to the other skewed distribution such as gamma distribution. From the density function for the lognormal distribution

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi x}} e^{-\frac{1}{2}(\frac{\log x - \mu}{\sigma})^2} \tag{2.24}$$

where $\mu \in (-\infty, +\infty)$ and $\sigma > 0$. And the maximum likelihood estimators for μ and σ^2 are

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \log x_i}{n} \tag{2.25}$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} (\log x_i - \mu)^2}{n} \tag{2.26}$$

Chapter 3

Results

3.1 Glycan distance comparison for different databases

We first compared the difference between each target and decoy glycan pair because we assumed that if a target glycan and the corresponding decoy glycan are too similar, especially if the fragment ions of *B*-ions and *Y*-ions are too similar, the software tool tends to assign decoy glycans to a spectrum. Since glycans have a tree structure, in which each monosaccharide can be treated as a node and each glycosidic bond can be treated as an edge, we calculated the difference metrics between a target glycan and the corresponding decoy glycan on the basis of tree edit distance.

There are three types of tree edit operations: insert operation inserts a node, delete operation deletes a node, and change operation relabels one node to another. Our decoy construction methods changed monosaccharides while maintaining the tree topologies, the distance can be calculated by checking each monosaccharide at a certain position in a tree. For example, if the root node in a target glycan tree is HexNAc and the same position of the node (i.e, the root node) in the corresponding decoy glycan tree is Hex, we count one to calculate the distance. Take Figure 3.1 as an example, the distance between target and decoy glycan is 4 and the ratio of different monosaccharides in a decoy glycan is 0.8.

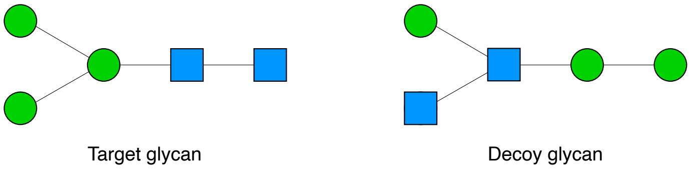


Figure 3.1: Permutation method

The average tree distance and the ratio of different monosaccharides in a decoy glycan tree for each decoy glycan construction method are shown in Table 3.1. The more sets of permutations were carried out, the more tree distance they had among the decoy construction by permutation methods. The swap methods had the ratio of different monosaccharides according to the swapping percentage. A random decoy database had enough distance between targets and decoys.

Table 3.1: Glycan distance comparison

Decoy construction	Average tree distance	Ratio of different monosaccharide
Permutation 10 sets	13.20	0.82
Permutation 30 sets	13.96	0.87
Permutation 60 sets	14.32	0.89
Swap 50%	8.23	0.52
Swap 75%	12.35	0.78
Swap 100%	15.97	1.00
Random	10.77	0.67

3.2 FDR estimation for different databases

We then calculated glycan FDR for each decoy glycan construction method using the assignment of glycans to all the spectra. From the largest glycan score to the smallest glycan score, we applied Equation 2.1 to all the GPSMs. After the glycan FDR calculation, we observed the number of GPSM at 1%, 2%, and 5%, and the threshold glycan scores for each FDR criteria (Table 3.2 to 3.4).

Also, we plotted the FDR curve to observe the relationships between a given FDR and the number of GPSM (Figure 3.2). As shown in the figure, permutation 10, permutation 30, and permutation 60 decoy databases had similar curves, although permutation 10 database had fewer number of matches between FDR of 2% and 4%. Permutation 30 and permutation 60 decoy databases had almost the same number of GPSM below the FDR of 2%, permutation 60 decoy database had more number of GPSM above FDR of 2%. Swap 50 % method had by far the fewest number of matches at 1% and 2% glycan FDR, while swap 75 % and swap 100 % methods had a much larger number of matches at any FDR threshold, for example, 3264 matches for swap 75 % method and 4672 matches for swap 100 % method at 1% FDR compared to 674 matches for permutation 60 method. Figure 3.3 to Figure 3.9 show target glycan score distribution and decoy glycan score distribution for each decoy database method, and Table 3.5 describes the number of target and decoy matches for each method. If you look at this information, Swap 50%, Swap 75%, and Swap 100% methods had significantly fewer decoy matches than the other methods, thus they had much larger GPSM. Therefore we cannot say that this database method is a reliable method for FDR estimation. We focused on permutation 10 sets, 30 sets, 60 sets, and random methods in the following analysis.

Table 3.2: Decoy method comparison at 1% glycan FDR

Decoy construction	Number of GPSM for 1% glycan FDR	Threshold glycan score
Permutation 10 sets	453	6.79
Permutation 30 sets	522	6.64
Permutation 60 sets	674	6.29
Swap 50%	65	8.81
Swap 75%	3264	0.65
Swap 100%	4672	0.14
Random	1072	5.25

Table 3.3: Decoy method comparison at 2% glycan FDR

Decoy construction	Number of GPSM for 2% glycan FDR	Threshold glycan score
Permutation 10 sets	877	5.77
Permutation 30 sets	1692	3.61
Permutation 60 sets	1884	3.00
Swap 50%	76	8.56
Swap 75%	4667	0.15
Swap 100%	5172	0.11
Random	1768	3.00

Table 3.4: Decoy method comparison at 5% glycan FDR

Decoy construction	Number of GPSM for 5% glycan FDR	Threshold glycan score
Permutation 10 sets	2673	1.40
Permutation 30 sets	2834	1.17
Permutation 60 sets	2967	0.99
Swap 50%	5050	0.12
Swap 75%	6070	0.08
Swap 100%	6491	0.07
Random	2651	1.46

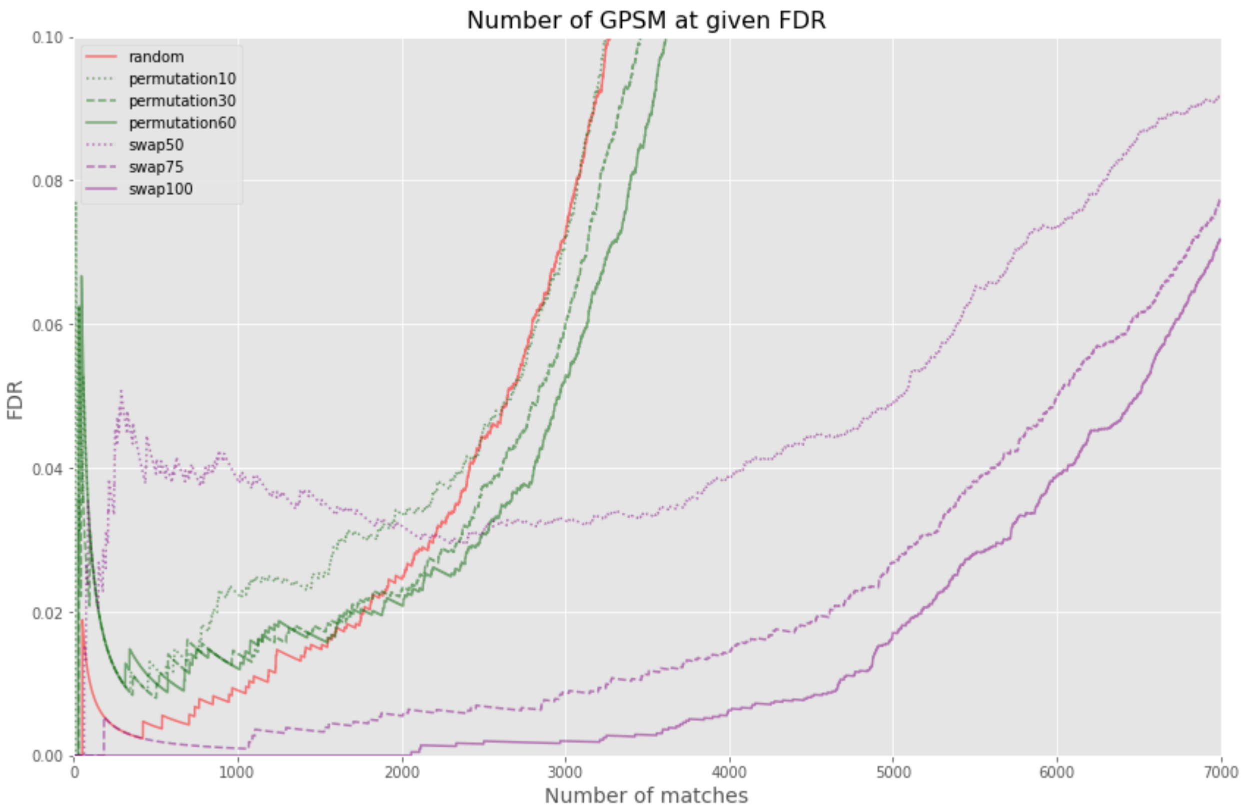


Figure 3.2: Number of GPSM at given glycan FDR

Table 3.5: Number of target/decoy matches for 100% glycan FDR

Decoy construction	Target matches	Decoy matches
Permutation 10 sets	4525	1449
Permutation 30 sets	4625	1281
Permutation 60 sets	4640	1244
Swap 50%	5172	338
Swap 75%	5273	196
Swap 100%	5288	140
Random	4642	1436

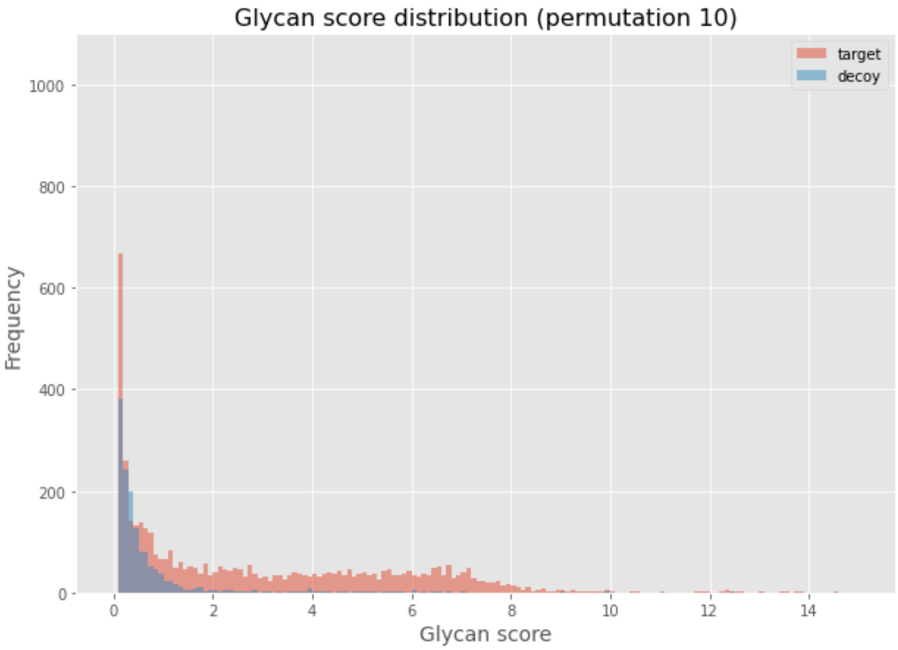


Figure 3.3: Glycan score distribution of permutaion 10 database

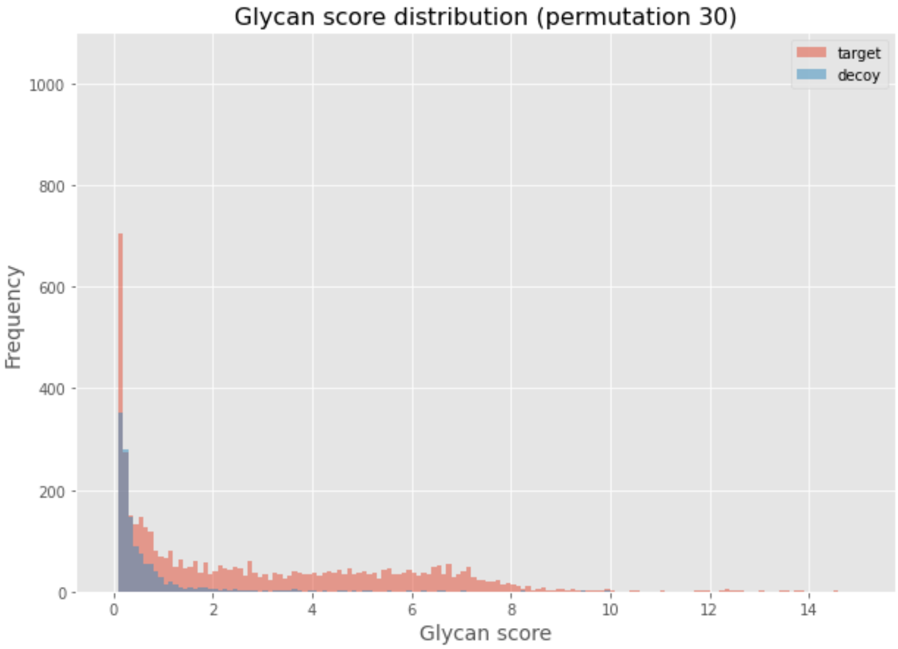


Figure 3.4: Glycan score distribution of permutaion 30 database

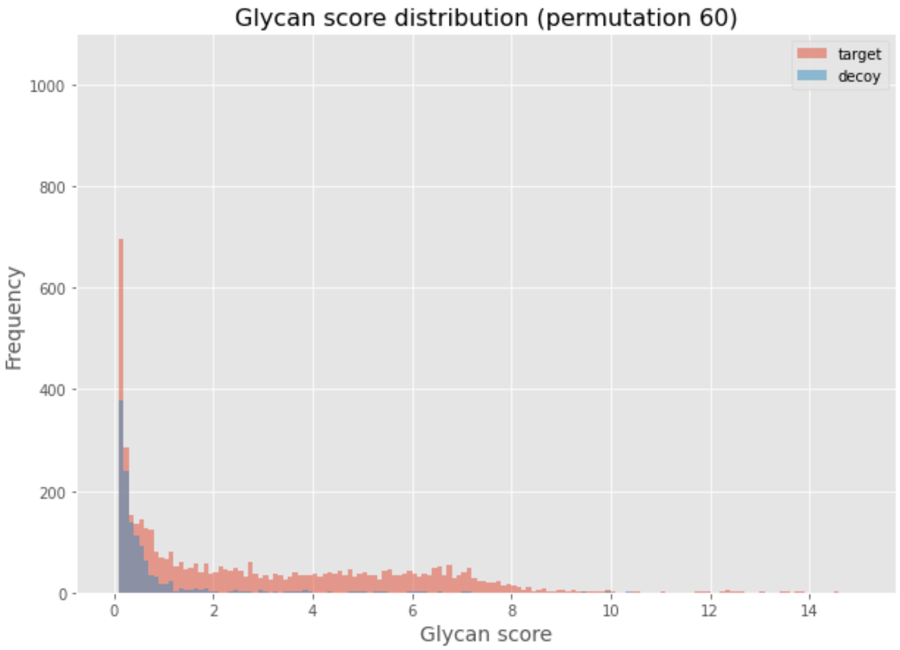


Figure 3.5: Glycan score distribution of permutaion 60 database

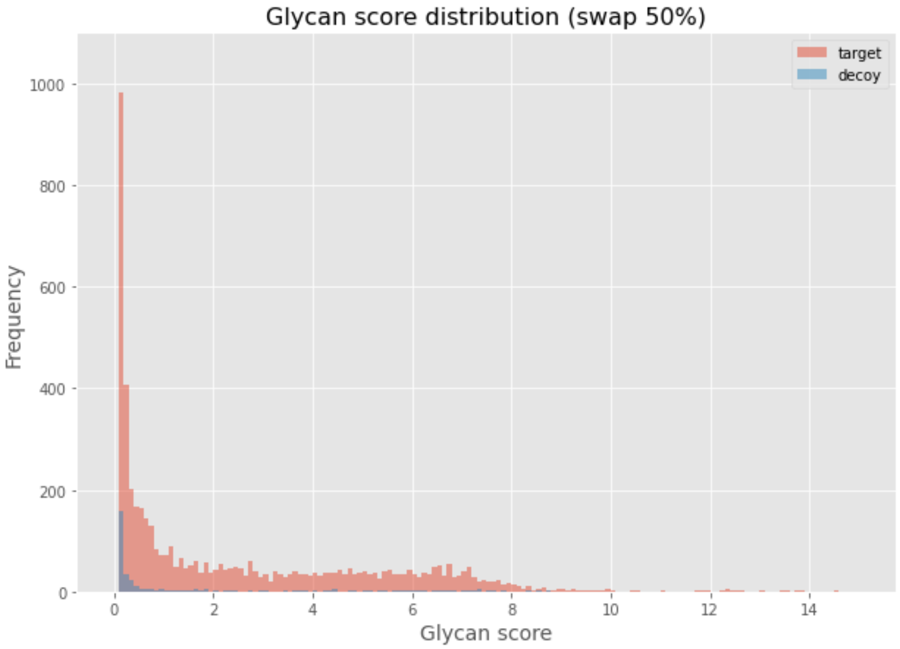


Figure 3.6: Glycan score distribution of swap 50% database

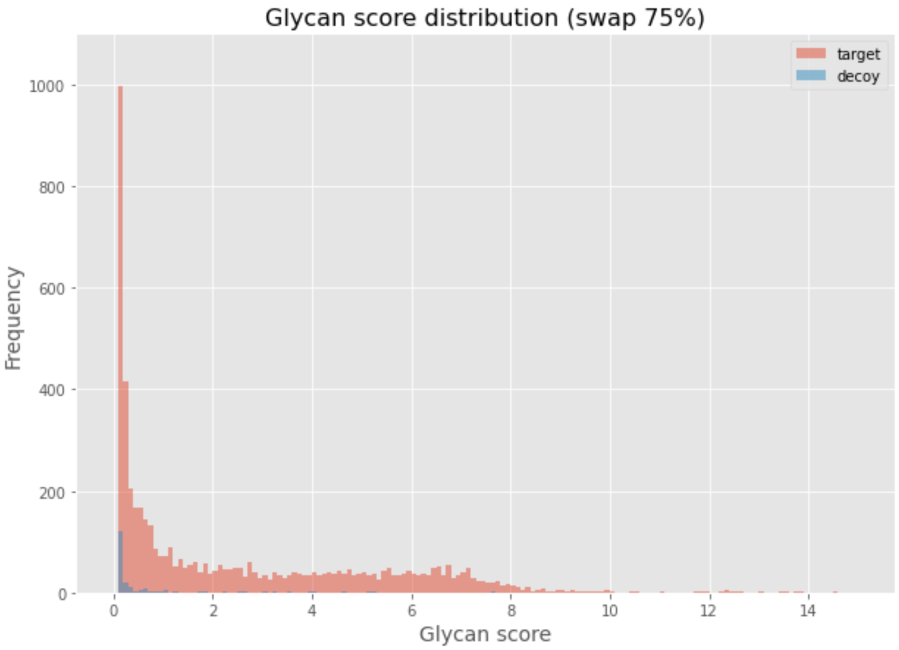


Figure 3.7: Glycan score distribution of swap 75% database

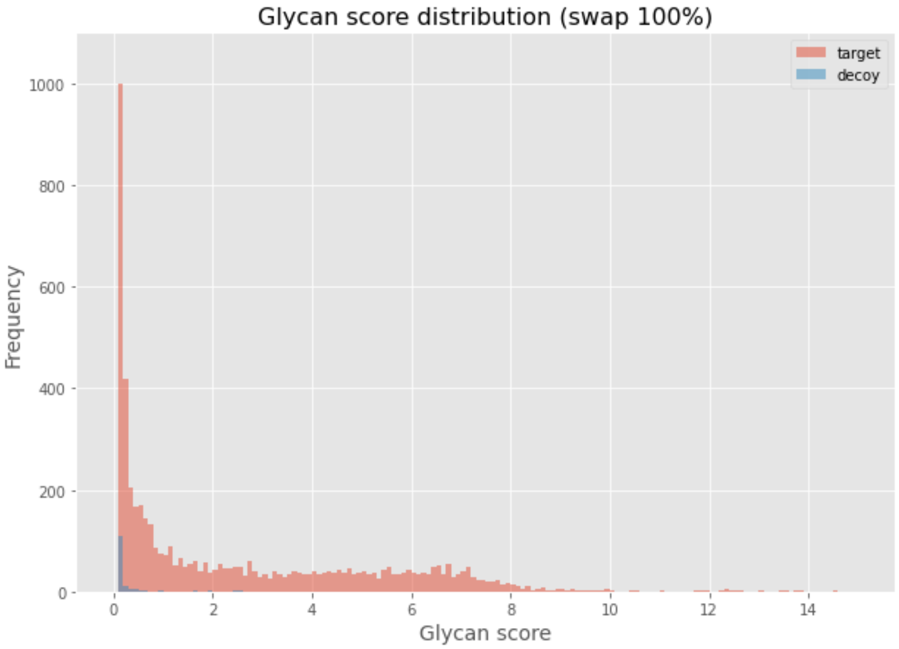


Figure 3.8: Glycan score distribution of swap 100% database

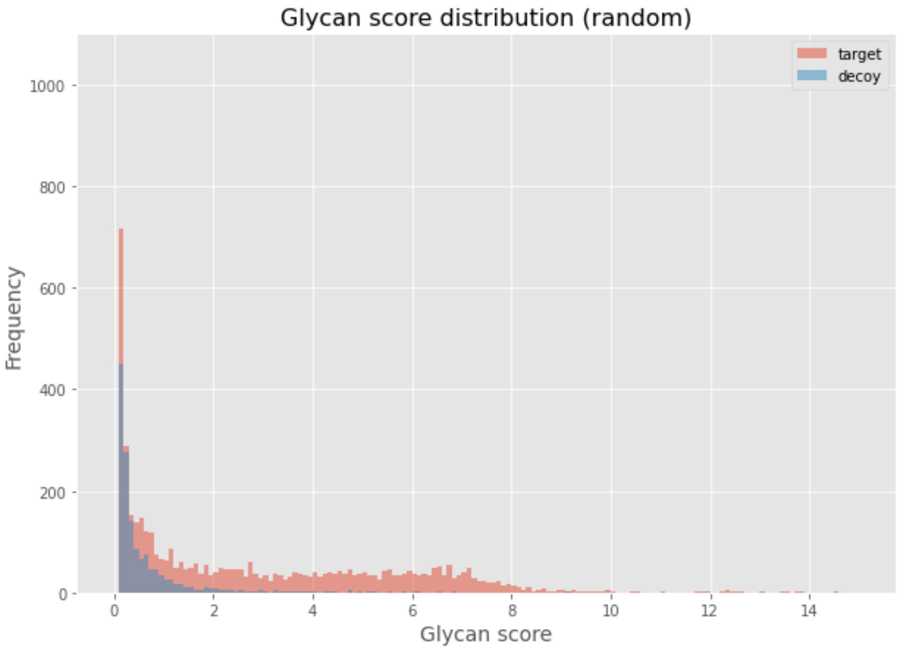


Figure 3.9: Glycan score distribution of random database

3.3 Mixture model-based comparison between incorrect target distribution and decoy distribution

The key assumption of the target-decoy search strategy is that false positive identifications are equally likely to come from the target database and the decoy database. To be consistent with this assumption, we first decomposed the target matches into target correct matches and target incorrect matches, which were normal distribution and lognormal distribution, respectively. In addition, we fitted the decoy glycan score histogram to lognormal distribution to compare it with the target incorrect match distribution.

Figure 3.10, 3.12, 3.14, and 3.16 show the histograms of target matches and estimated components of incorrect target score distributions for each decoy database method. Figure 3.11, 3.13, 3.15, and 3.17 show the histograms of decoy matches and fitted decoy score distributions for each decoy database method. In order to precisely compare the shape of target incorrect distribution and decoy distribution, we plotted cumulative density for each distribution in Figure 3.18 to 3.21. If the target cumulative distribution curve and decoy cumulative distribution curve are completely overlapped, we can assume that the two distributions are identical and have the same parameters. On the other hand, if the two curves are plotted further from each other, the shapes of these distributions are not so similar, which means the decoy database method is not consistent with the assumption of the target-decoy search strategy.

At a first glance, permutation 10, permutation 30, and permutation 60 methods had less gap between the two distribution curves, whereas the random method had a wider gap between the curves, which means the random method did not have similar target incorrect distribution and decoy distribution. More closely looking at Figure 3.18 to 3.21, we compared permutation 10, permutation 30, and permutation 60 methods in detail, especially the cumulative distribution function at a threshold glycan score of 1% FDR. The threshold score for 1% glycan FDR score for the permutation 10 method was 6.79. The cumulative distribution function of 6.79 in target incorrect distribution was 0.982 and the cumulative distribution function of 6.79 in decoy distribution was 0.990. The threshold score for 1% glycan FDR score for the permutation 30 method was 6.64. The cumulative distribution function of 6.64 in target incorrect distribution was 0.983 and the cumulative distribution function of 6.64 in decoy distribution was 0.991. The threshold score for 1% glycan FDR score for the permutation 60 method was 6.29. The cumulative distribution function of 6.29 in target incorrect distribution was 0.981 and the cumulative distribution function of 6.29 in decoy distribution was 0.991.

From this analysis, the gap between the cumulative distribution function at a threshold score of 1% FDR for target incorrect distribution and decoy distribution observed by the permutation 60 method was larger and the gaps observed by permutation 10 and permutation 30 were smaller compared to each other. Taking account of the assumption in the target-decoy strategy that there is an equal probability of incorrect matches from the target database and the decoy database, the permutation 10 method and the permutation 30 method are more suitable decoy databases with regard to the similarity between target incorrect distribution and decoy distribution at their threshold score for 1% FDR.

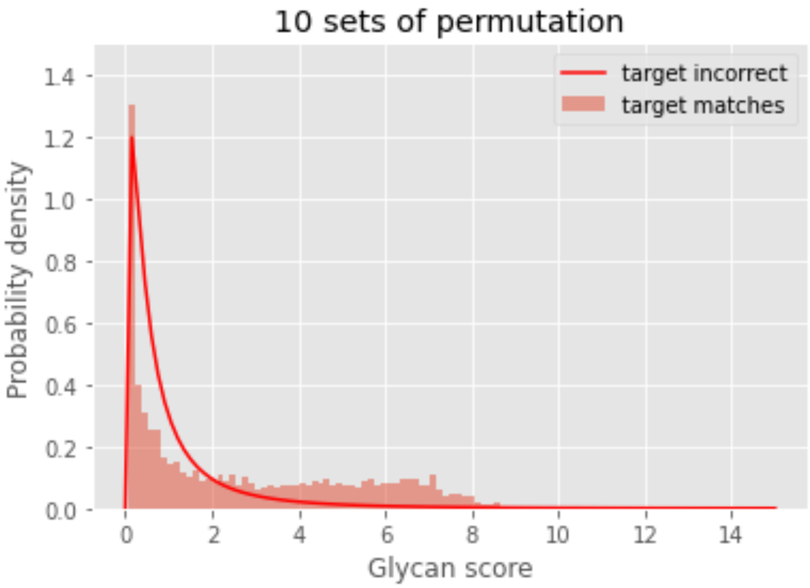


Figure 3.10: Estimated target incorrect glycan distribution of 10 sets of permutation database

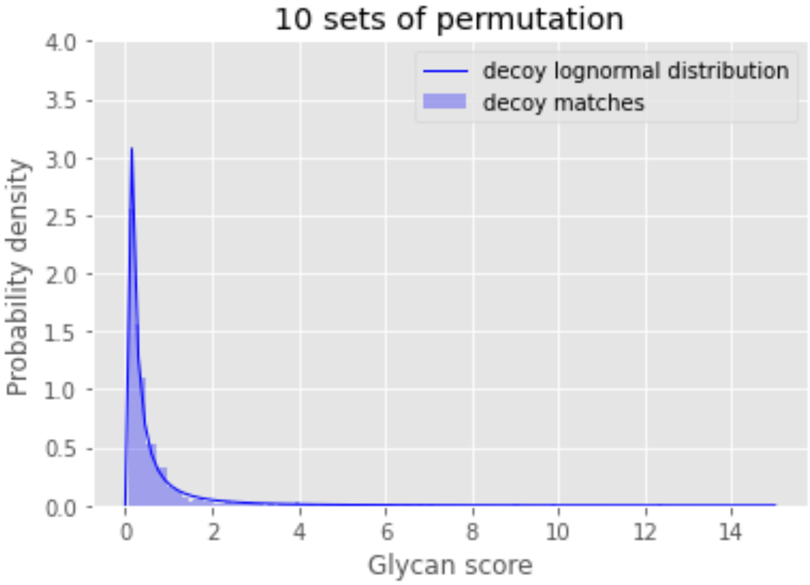


Figure 3.11: Fitting glycan score distribution of 10 sets of permutation database

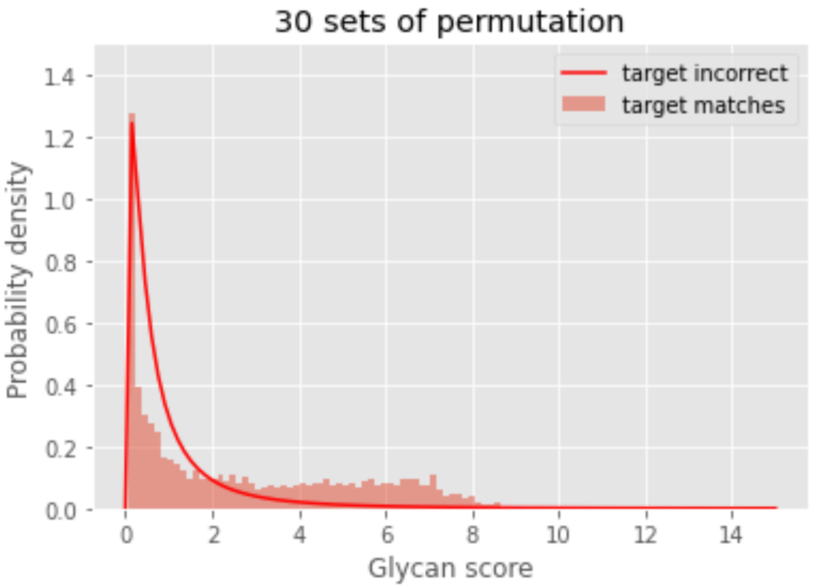


Figure 3.12: Estimated target incorrect glycan distribution of 30 sets of permutation database

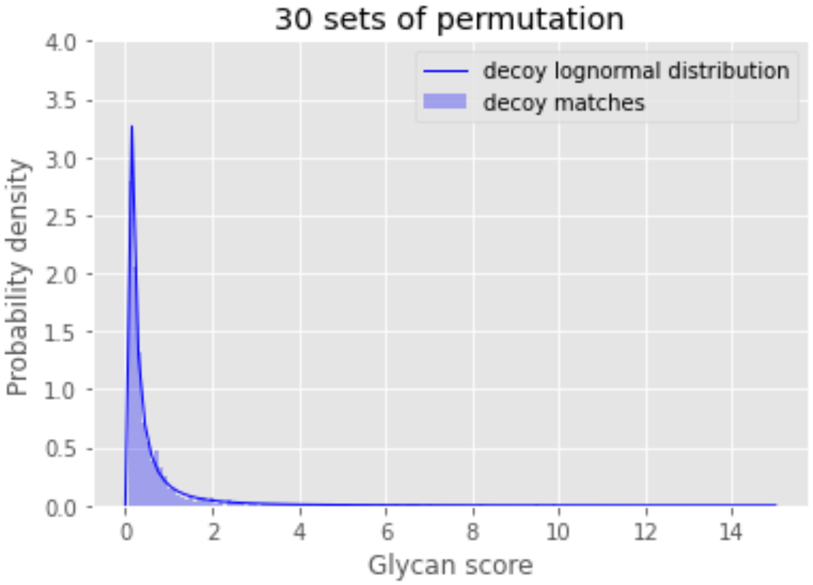


Figure 3.13: Fitting glycan score distribution of 30 sets of permutation database

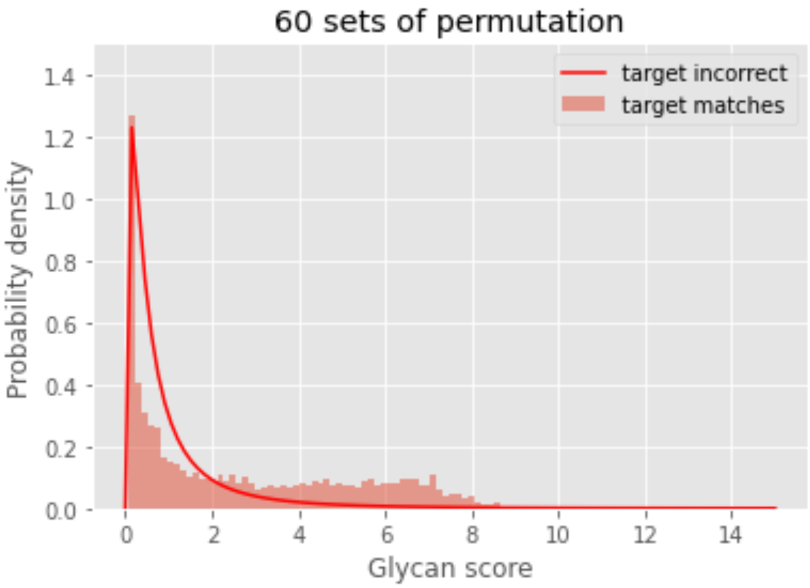


Figure 3.14: Estimated target incorrect glycan distribution of 60 sets of permutation database

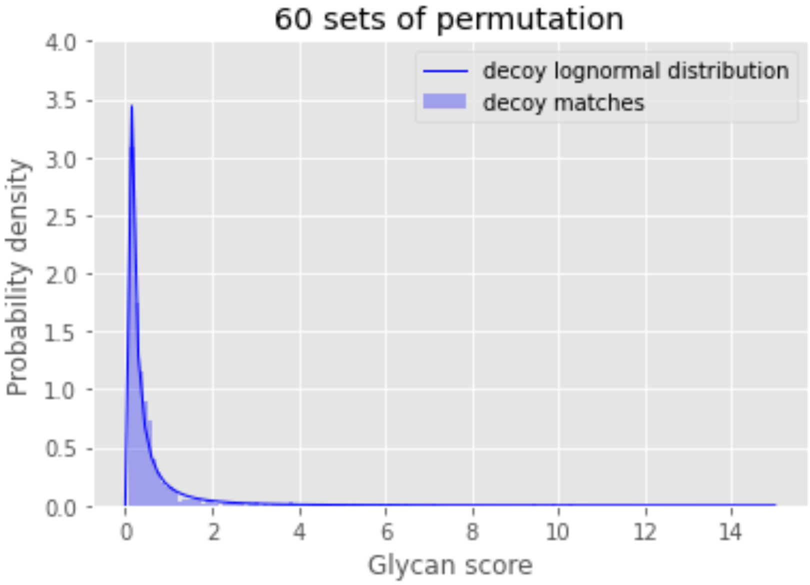


Figure 3.15: Fitting glycan score distribution of 60 sets of permutation database

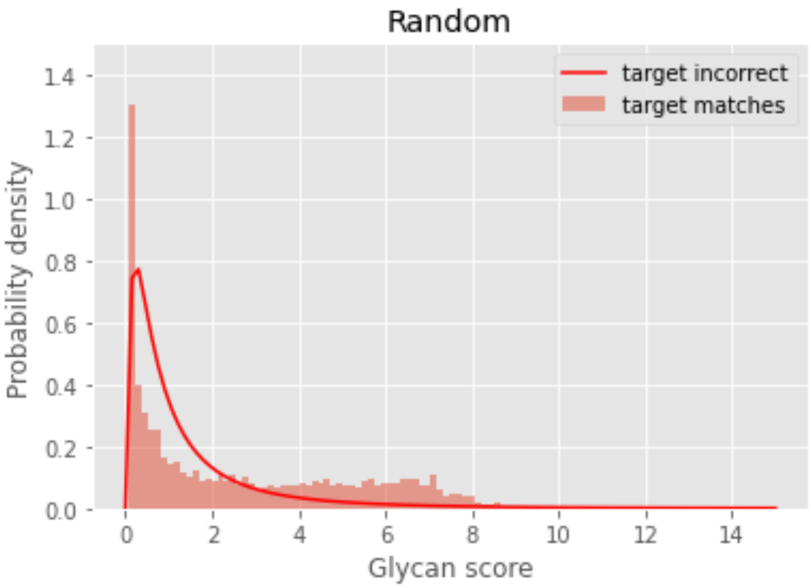


Figure 3.16: Estimated target incorrect glycan distribution of random database

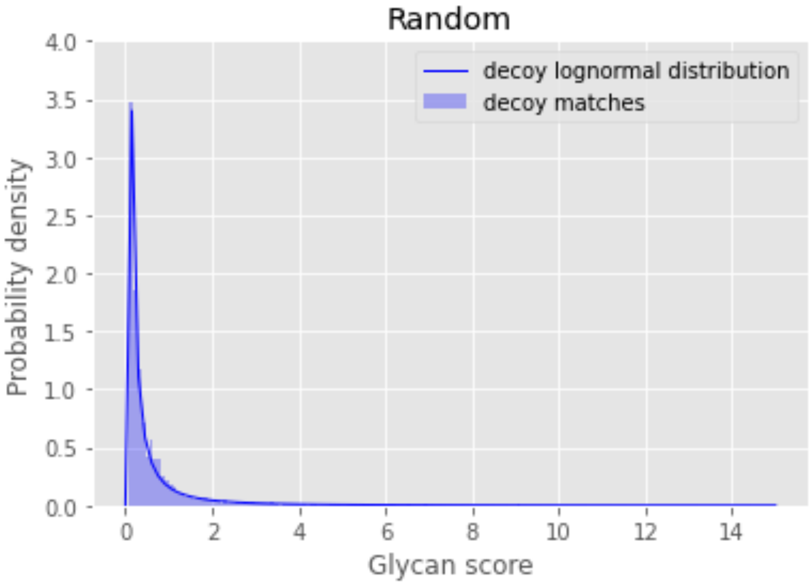


Figure 3.17: Fitting glycan score distribution of random database

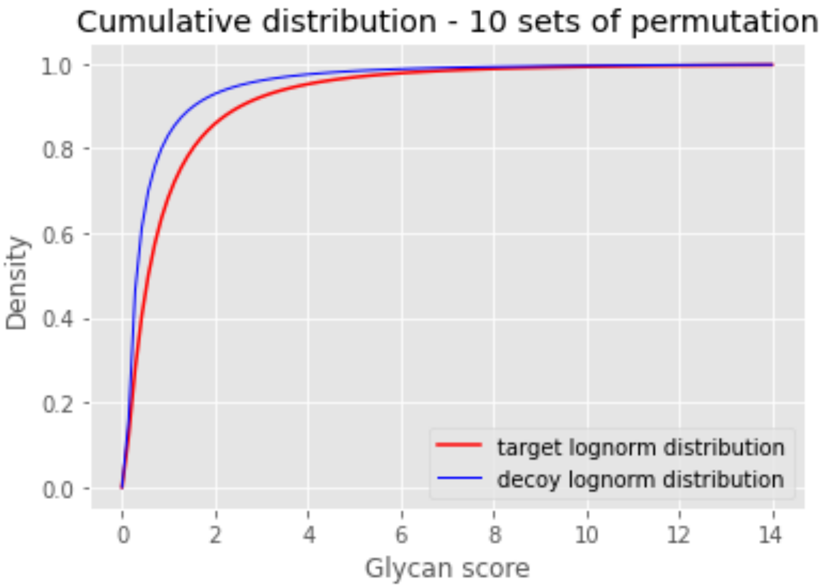


Figure 3.18: Cumulative density of incorrect target distribution and decoy distribution for permutation 10 database

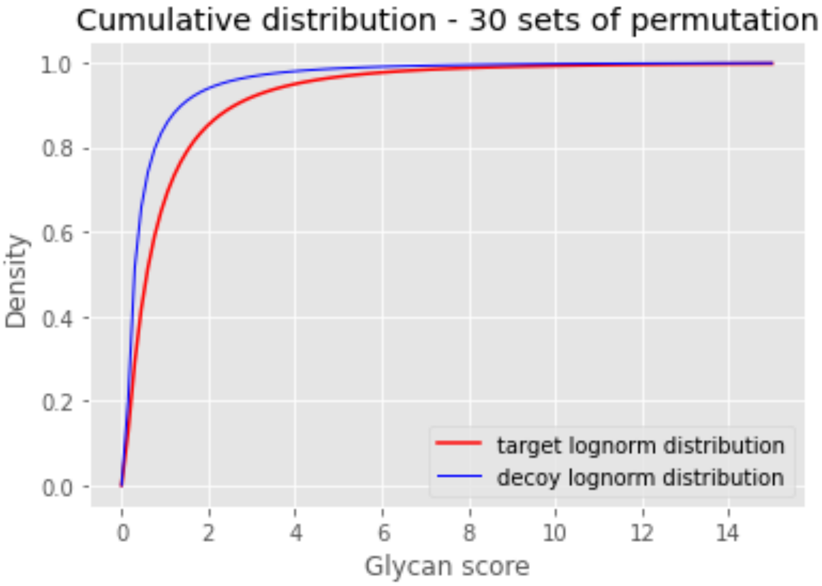


Figure 3.19: Cumulative density of incorrect target distribution and decoy distribution for permutation 30 database

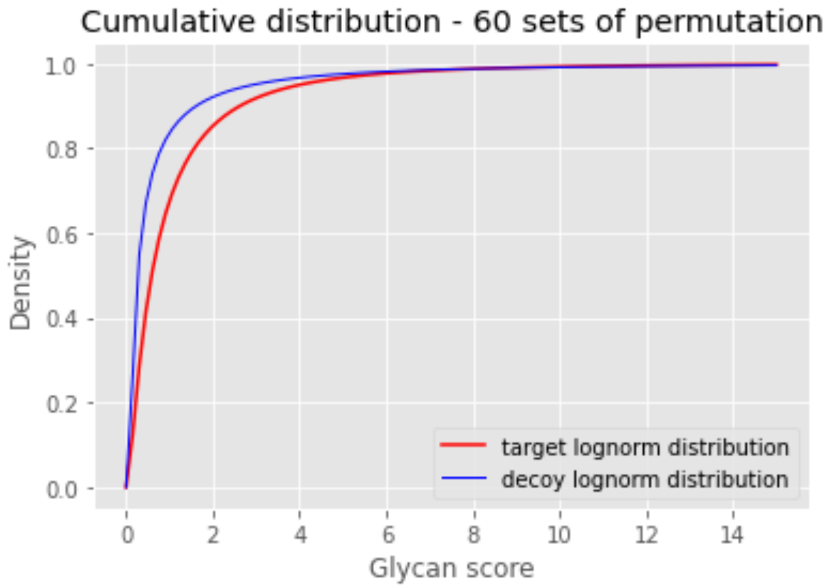


Figure 3.20: Cumulative density of incorrect target distribution and decoy distribution for permutation 60 database

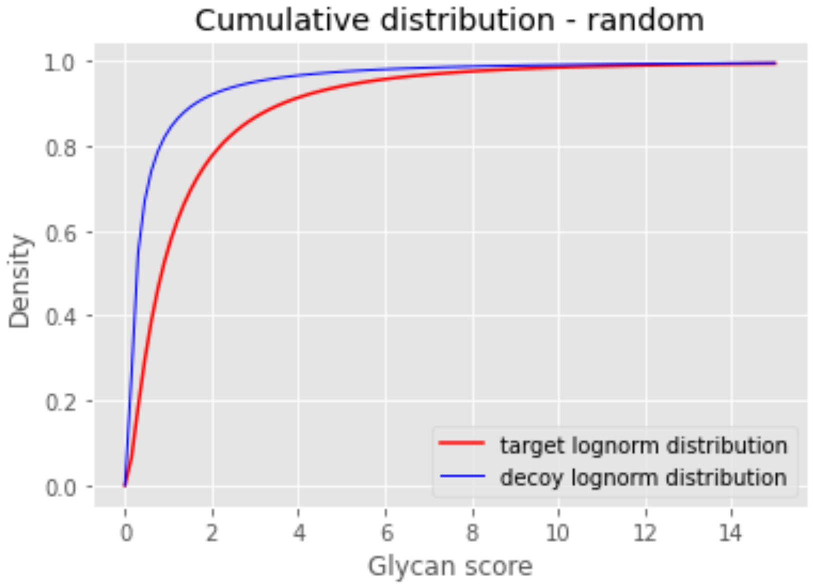


Figure 3.21: Cumulative density of incorrect target distribution and decoy distribution for random database

3.4 Overlapped identifications between GlycanFinder and pGlyco

To evaluate the performance of each decoy construction method, we counted overlapped glycopeptide identifications by using another glycopeptide sequencing software. We used pGlyco 2.0 as a benchmark software because it is currently one of the most widely used software tools. GlycanFinder employs its default decoy glycan method, which adds random mass to the theoretical target glycan mass list to generate decoy spectra, and the software searches experimental spectra against target and decoy glycan spectra. The new version of GlycanFinder has been developed as of now, but we compare the results obtained from the previous version of GlycanFinder in this thesis.

Table 3.6 summarizes the overlapped identifications between GlycanFinder and pGlyco 2.0 for 1% glycan FDR. The number of overlapped identification for swap 75 % and swap 100 % were larger than other methods. However, considering the ratio of the number of overlapped identifications between GlycanFinder and pGlyco 2.0 to the number of identifications in GlycanFinder, the ratios for swap methods were smaller than that for permutation methods. Thus, a certain portion of the GlycanFinder identifications by swap 75% and swap 100% method seems to be not so reliable, although these decoy methods achieved a larger number of identifications for 1% FDR.

Table 3.6: Number of overlapped identifications between GlycanFinder and pGlyco 2.0

Decoy construction	Number of identifications in GlycanFinder	Overlapped identifications with pGlyco 2.0
Default	349	74
Permutation 10 sets	453	86
Permutation 30 sets	522	96
Permutation 60 sets	674	119
Swap 50%	65	8
Swap 75%	3264	264
Swap 100%	4672	269
Random	1072	151

Chapter 4

Conclusion and discussion

In this study, we constructed various kinds of decoy glycan databases on the basis of the target-decoy search strategy: the permutation method, the swap method, and the random method. To evaluate the effectiveness of each glycan construction method, we compared these databases from the viewpoint of the tree edit distance between the decoy glycan sequences and the corresponding target glycans, the number of GPSM at a given FDR threshold, and the distribution of glycan target scores and decoy scores for 100% glycan FDR. Also, we observed the number of overlapped glycopeptide identifications between GlycanFinder and a benchmark glycopeptide identification software pGlyco 2.0. Furthermore, we estimated a distinction between the correct assignments and the incorrect assignments among the target glycan score distribution using the MCMC-based mixture model and compared the shape of the target incorrect distributions and the decoy glycan score distributions for each decoy construction method following the key assumption of the target-decoy strategy.

From the experimental results, we concluded that the decoy glycan database generated by the permutation method is a more reasonable method than the swap method and the random method for FDR estimation of glycopeptide identification based on the observation of the number of GPSM and the ratio of the number of GlycanFinder identifications to the number of overlapped identifications with pGlyco 2.0. With regard to the comparison of the shape of the target incorrect distributions and the decoy glycan score distributions, the permutation approaches achieved a more similar shape of the two distributions, which can satisfy the assumption of equal likeliness of false positives in the target database and the decoy database in the target-decoy search strategy. Among the permutation methods, the gap of the cumulative distribution function of a threshold glycan score of 1% FDR between the target incorrect distribution and the decoy distribution was larger in the permutation 60 method, whereas the gap was smaller in the permutation 10 and the permutation 30 method. However, the difference in the metrics among these methods is rather small and the difference may have come from the difference in the threshold score of 1% FDR.

There are, however, some possible limitations in this study. In this work, we used the mouse brain glycopeptide dataset for FDR estimation of glycopeptide identification yielded by GlycanFinder, and compared these results against those produced by the benchmark software pGlyco 2.0. More various kinds of datasets can be used to validate the performance of decoy glycan construction approaches, and another benchmark glycopeptide identification software apart from pGlyco 2.0 for a better comparison of the number of their identifications in future work.

Another limitation of this study concerns the mixture model to distinguish correct assignments and incorrect assignments in the target score distribution. We employed the unsupervised normal-lognormal mixture model and compared the distribution curves. A further study could employ a semisupervised mixture model using the information from decoy matches, which represents incorrect assignments. Also, another family of distributions apart from normal and lognormal distribution could be considered for modeling the two-component mixture.

Furthermore, we constructed decoy glycan databases by assigning monosaccharides to each tree node while maintaining the glycan tree structure in this study to have a similar monosaccharide composition as the glycans in the target database, specifically for the permutation method and the random method. In future investigations, it might be possible to construct a decoy glycan database by changing the glycan tree structure from the corresponding target tree structure.

Bibliography

[1] Rolf Apweiler, Henning Hermjakob, and Nathan Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database11Dedicated to Prof. Akira Kobata and Prof. Harry Schachter on the occasion of their 65th birthdays. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1473(1):4–8, December 1999.

[2] Hyun Joo An, John W Froehlich, and Carlito B Lebrilla. Determination of glycosylation sites and site-specific heterogeneity in glycoproteins. *Current Opinion in Chemical Biology*, 13(4):421–426, October 2009.

[3] Kristel Kodar, Johannes Stadlmann, Kersti Klaamas, Boris Sergeyev, and Oleg Kurtenkov. Immunoglobulin G Fc N-glycan profiling in patients with gastric cancer by LC-ESI-MS: relation to tumor progression and survival. *Glycoconjugate Journal*, 29(1):57–66, January 2012.

[4] Philippe Van den Steen, Pauline M. Rudd, Raymond A. Dwek, and Ghislain Opdenakker. Concepts and Principles of O-Linked Glycosylation. *Critical Reviews in Biochemistry and Molecular Biology*, 33(3):151–208, January 1998.

[5] Allison Doerr. Glycoproteomics. *Nature Methods*, 9(1):36–36, January 2012.

[6] Richard D. Cummings. The repertoire of glycan determinants in the human glycome. *Molecular BioSystems*, 5(10):1087–1104, 2009. Publisher: Royal Society of Chemistry.

[7] Scott Doubet, Klaus Bock, Dana Smith, Alan Darvill, and Peter Albersheim. The complex carbohydrate structure database. *Trends in Biochemical Sciences*, 14(12):475–477, December 1989.

[8] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl_1):D277–D280, January 2004.

[9] Satya S. Sahoo, Christopher Thomas, Amit Sheth, Cory Henson, and William S. York. GLYDE—an expressive XML standard for the representation of glycan structure. *Carbohydrate Research*, 340(18):2802–2807, December 2005.

[10] N. Kikuchi, A. Kameyama, S. Nakaya, H. Ito, T. Sato, T. Shikanai, Y. Takahashi, and H. Narimatsu. The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics*, 21(8):1717–1718, April 2005.

[11] Ehud Banin, Yael Neuberger, Yaniv Altshuler, Asaf Halevi, Ori Inbar, Dotan Nir, Avinoam Dukler, and author_in_Japanese. A Novel Linear Code® Nomenclature for Complex Carbohydrates. *Trends in Glycoscience and Glycotechnology*, 14(77):127–137, 2002.

[12] Andreas Bohne-Lang, Elke Lang, Thomas Förster, and Claus-W. von der Lieth. LINUCS: LInear Notation for Unique description of Carbohydrate Sequences. *Carbohydrate Research*, 336(1):1–11, November 2001.

[13] Stephan Herget, Philip V. Toukach, René Ranzinger, William E. Hull, Yuriy A. Knirel, and Claus-Wilhelm von der Lieth. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Structural Biology*, 8(1):35, August 2008.

[14] S. Herget, R. Ranzinger, K. Maass, and C. W. v. d. Lieth. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydrate Research*, 343(12):2162–2171, August 2008.

[15] Alessio Ceroni, Anne Dell, and Stuart M. Haslam. The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code for Biology and Medicine*, 2(1):3, August 2007.

[16] Anca-Narcisa Neagu, Madhuri Jayathirtha, Emma Baxter, Mary Donnelly, Brindusa Alina Petre, and Costel C. Darie. Applications of Tandem Mass Spec-

trometry (MS/MS) in Protein Analysis for Biomedical Research. *Molecules*, 27(8):2411, January 2022. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[17] T H Plummer, J H Elder, S Alexander, A W Phelan, and A L Tarentino. Demonstration of peptide:N-glycosidase F activity in endo-beta-N-acetylglucosaminidase F preparations. *Journal of Biological Chemistry*, 259(17):10700–10704, September 1984.

[18] Li Cao, Nikola Tolić, Yi Qu, Da Meng, Rui Zhao, Qibin Zhang, Ronald J. Moore, Erika M. Zink, Mary S. Lipton, Ljiljana Paša-Tolić, and Si Wu. Characterization of intact N- and O-linked glycopeptides using higher energy collisional dissociation. *Analytical Biochemistry*, 452:96–102, May 2014.

[19] J. Mitchell Wells and Scott A. McLuckey. Collision ‐ Induced Dissociation (CID) of Peptides and Proteins. In *Methods in Enzymology*, volume 402 of *Biological Mass Spectrometry*, pages 148–185. Academic Press, January 2005.

[20] Klaus Biemann. [25] Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. In *Methods in Enzymology*, volume 193 of *Mass Spectrometry*, pages 455–479. Academic Press, January 1990.

[21] James A. Madsen, Victor Farutin, Yin Yin Lin, Stephen Smith, and Ishan Capila. Data-independent oxonium ion profiling of multi-glycosylated biotherapeutics. *mAbs*, 10(7):968–978, August 2018.

[22] Roman A Zubarev. Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology*, 15(1):12–16, February 2004.

[23] John E. P. Syka, Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences*, 101(26):9528–9533, June 2004. Publisher: Proceedings of the National Academy of Sciences.

[24] Christopher Hughes, Bin Ma, and Gilles A. Lajoie. De Novo Sequencing Methods in Proteomics. In Simon J. Hubbard and Andrew R. Jones, editors, *Proteome Bioinformatics*,

volume 604, pages 105–121. Humana Press, Totowa, NJ, 2010. Series Title: Methods in Molecular Biology.

[25] Hyun Joo An and Carlito B. Lebrilla. Structure elucidation of native N- and O-linked glycans by tandem mass spectrometry (tutorial). *Mass Spectrometry Reviews*, 30(4):560–578, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mas.20283.

[26] Pauline M. Rudd, Niclas G. Karlsson, Kay-Hooi Khoo, Morten Thaysen-Andersen, Lance Wells, and Nicolle H. Packer. Glycomics and Glycoproteomics. In Ajit Varki, Richard D. Cummings, Jeffrey D. Esko, Pamela Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 4th edition, 2022.

[27] Hannes Hinneburg, Kathrin Stavenhagen, Ulrike Schweiger-Hufnagel, Stuart Pengelley, Wolfgang Jabs, Peter H. Seeberger, Daniel Varón Silva, Manfred Wuhrer, and Daniel Kolarich. The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. *Journal of the American Society for Mass Spectrometry*, 27:507–519, 2016.

[28] Sz-Wei Wu, Tsung-Hsien Pu, Rosa Viner, and Kay-Hooi Khoo. Novel LC-MS2 Product Dependent Parallel Data Acquisition Function and Data Analysis Workflow for Sequencing and Identification of Intact Glycopeptides. *Analytical Chemistry*, 86(11):5478–5486, June 2014. Publisher: American Chemical Society.

[29] Julian Saba, Sucharita Dutta, Eric Hemenway, and Rosa Viner. Increasing the Productivity of Glycopeptides Analysis by Using Higher-Energy Collision Dissociation-Accurate Mass-Product-Dependent Electron Transfer Dissociation. *International Journal of Proteomics*, 2012:1–7, May 2012.

[30] Qing Yu, Bowen Wang, Zhengwei Chen, Go Urabe, Matthew S. Glover, Xudong Shi, Lian-Wang Guo, K. Craig Kent, and Lingjun Li. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characteriza-

tion. *Journal of the American Society for Mass Spectrometry*, 28(9):1751–1764, September 2017. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[31] Kshitij Khatri, Yi Pu, Joshua A. Klein, Juan Wei, Catherine E. Costello, Cheng Lin, and Joseph Zaia. Comparison of Collisional and Electron-Based Dissociation Modes for Middle-Down Analysis of Multiply Glycosylated Peptides. *Journal of The American Society for Mass Spectrometry*, 29(6):1075–1085, June 2018.

[32] John R. Yates III. Database searching using mass spectrometry data. *ELECTROPHORESIS*, 19(6):893–900, 1998. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/elps.1150190604.

[33] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, January 2015.

[34] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(suppl_1):D61–D65, January 2007.

[35] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[36] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*, 20(18):3551–3567, 1999. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291522-2683%2819991201%2920%3A18%3C3551%3A%3AAID-ELPS3551%3E3.0.CO%3B2-2.

[37] Johannes Griss. Spectral library searching in proteomics. *PROTEOMICS*, 16(5):729–740, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201500296.

[38] Christopher Hughes, Bin Ma, and Gilles A. Lajoie. De Novo Sequencing Methods in Proteomics. *Proteome Bioinformatics*, pages 105–121, 2010. Publisher: Humana Press.

[39] Changjiang Xu and Bin Ma. Software for computational peptide identification from MS – MS data. *Drug Discovery Today*, 11(13):595–600, July 2006.

[40] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.1196.

[41] Bin Ma. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, November 2015. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[42] Ari Frank and Pavel Pevzner. PepNovo: ☒ De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry*, 77(4):964–973, February 2005. Publisher: American Chemical Society.

[43] Joshua E. Elias, Wilhelm Haas, Brendan K. Faherty, and Steven P. Gygi. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods*, 2(9):667–675, September 2005. Number: 9 Publisher: Nature Publishing Group.

[44] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, March 2007. Number: 3 Publisher: Nature Publishing Group.

[45] Suruchi Aggarwal and Amit Kumar Yadav. False Discovery Rate Estimation in Proteomics. In Klaus Jung, editor, *Statistical Analysis in Proteomics*, Methods in Molecular Biology, pages 119–128. Springer, New York, NY, 2016.

[46] Lukas Käll, John D. Storey, Michael J. MacCoss, and William Stafford Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, January 2008. Publisher: American Chemical Society.

[47] Juan Antonio Vizcaíno, Richard G. Côté, Attila Csordas, José A. Dianes, Antonio Fabregat, Joseph M. Foster, Johannes Griss, Emanuele Alpi, Melih Birim, Javier Contell, Gavin O' Kelly, Andreas Schoenegger, David Ovelleiro, Yasset Pérez-Riverol, Florian Reisinger, Daniel Ríos, Rui Wang, and Henning Hermjakob. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, 41(D1):D1063–D1069, January 2013.

[48] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J. Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. In David Edwards, editor, *Plant Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, pages 23–54. Springer, New York, NY, 2016.

[49] Kaizhong Zhang. A constrained edit distance between unordered labeled trees. *Algorithmica*, 15(3):205–222, March 1996.

[50] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A. Lajoie, and Bin Ma. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics : MCP*, 11(4):M111.010587, April 2012.

[51] Ming-Qi Liu, Wen-Feng Zeng, Pan Fang, Wei-Qian Cao, Chao Liu, Guo-Quan Yan, Yang Zhang, Chao Peng, Jian-Qiang Wu, Xiao-Jin Zhang, Hui-Jun Tu, Hao Chi, Rui-Xiang Sun, Yong Cao, Meng-Qiu Dong, Bi-Yun Jiang, Jiang-Ming Huang, Hua-Li Shen,

Catherine C. L. Wong, Si-Min He, and Peng-Yuan Yang. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nature Communications*, 8(1):438, September 2017. Number: 1 Publisher: Nature Publishing Group.

[52] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. Publisher: American Institute of Physics.

[53] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.