2011-01-01

Computational Tool For Automated Large-Scale Gpiomic Analysis

Juan Clemente Aguilar
University of Texas at El Paso, clemente.aguilar@gmail.com

COMPUTATIONAL TOOL FOR AUTOMATED LARGE-SCALE GPIOMIC ANALYSIS

JUAN CLEMENTE AGUILAR BONAVIDES

DEPARTMENT OF MATHEMATICAL SCIENCES

APPROVED:

_____

Ming-Ying Leung, Ph.D., Chair

_____

Igor C. Almeida, Ph.D., Co-Chair

_____

Tunna Baruah, Ph.D.

_____

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

Dedication

I dedicate this thesis to my wife, Ruth, for her unconditional support and love. Also to Oliver and Clara, who were always willing to participate in my study breaks.

COMPUTATIONAL TOOL FOR AUTOMATED LARGE-SCALE GPIOMIC ANALYSIS

by

JUAN CLEMENTE AGUILAR BONAVIDES, M. Sc.

THESIS

PRESENTED TO THE FACULTY OF THE GRADUATE SCHOOL OF

THE UNIVERSITY OF TEXAS AT EL PASO

IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICAL SCIENCES

THE UNIVERSITY OF TEXAS AT EL PASO

MAY 2011

Acknowledgements

I owe my deepest gratitude to my mentors Dr. Ming-Ying Leung and Dr. Igor C. Almeida for their encouragement, patience and guidance.

I am heartily thankful to my collaborator and friend, Dr. Ernesto Nakayasu, whose enthusiasm and support enabled me to develop an understanding of the subject.

I also want to acknowledge Dr. Tunna Baruah for kindly accepting being a part of my thesis committee; Dr. Emma Arigi, Mr. Leonel Saldivar, Mr. Felipe G. Lopes, Ms. Paula Gonzalez Parra, Mr. Anibal Sosa, Mr. Carlos Ramirez, Mr. Julio C. Olaya and Mr. Reinaldo Sanchez-Arias for their productive comments and sharing their knowledge.

This work was supported by NIH grants R01AI070655, 3R01AI070655-04S1, 2G12RR008124-16A1, and 2G12RR008124-16A1S1; NHARP grant 003661-0013-2007; and NSF grant DMS0800272.

Abstract

Liquid chromatography-tandem mass spectrometry (LC-MS/MS or MS/MS) is the most efficient tool today for the identification of glycosylphosphatidylinositol (GPI) molecules. The amount of data produced in each MS/MS experiment is a major bottleneck in high-throughput GPIomic (the entire collection of free and protein-linked GPIs) projects. Efficient computational tools can significantly reduce the amount of time analyzing MS/MS data; however, at present the automatic interpretation of these data to annotate GPI structures is absent. We propose a library-based tool to identify GPI structures by matching fragment peaks in the spectra with data derived from a theoretical database of GPI structures that we have developed. Currently, our scoring method produces an overlap between the scores of correct and incorrect structures identified since a number of isobaric structures exist within the database. Thus, to ensure that most of the identifications are true positive, a scoring system with better specificity must be applied. Considering the success of the peptide identification approach and the new methodologies for glycan and lipid identification, we expect that the development of a new method that combines techniques can be developed in combination with our existing tool.

Table of Contents

List of Tables

Chapter 3

Chapter 5

List of Figures

Chapter 1: Introduction

The cell membrane of living organisms contain embedded proteins, which are involved in a variety of cellular processes. Some of these proteins are anchored to the cell membrane by a glycolipid. The molecule glycosylphosphatidylinositol (abbreviated GPI-anchor or GPI), is a glycolipid composed of a polysaccharide (glycan) group and a lipid group. In mammalian cells and several protozoa some free GPIs are found at the cell surface. Free GPIs are sometimes referred to as GIPLs and share a common structure to protein-linked GPIs.

GPIs have a broad presence in living organisms; they have been identified in many eukaryotes, including humans, and are particularly abundant in protozoa. GPI-anchored proteins are involved in a number of functions such as enzymatic catalysis, adhesion, and in some cases they can mediate signal transduction across the plasma membrane (Ikezawa 2002). GPIs are important in medicine. Inhibitors of GPI biosynthesis could be drug targets for diseases caused by pathogenic protozoa (Nosjean et al., 1997).

The biosynthesis of GPI-anchored proteins is carried out in the Endoplasmic Reticulum (ER) (Ferguson et al., 1999). Depending on the protein to which they are attached and the organism in which they are synthesized GPI structures are very heterogeneous. A typical core structure of GPI is composed of a lipid group attached to a glycan group via inositol-phosphate; this glycolipid is then bounded to a mature protein (Figure 1.1). The lipid group allows the entire molecule to anchor the protein to the cell membrane. Heterogeneity in GPI anchors is derived from various substitutions of this core structure.

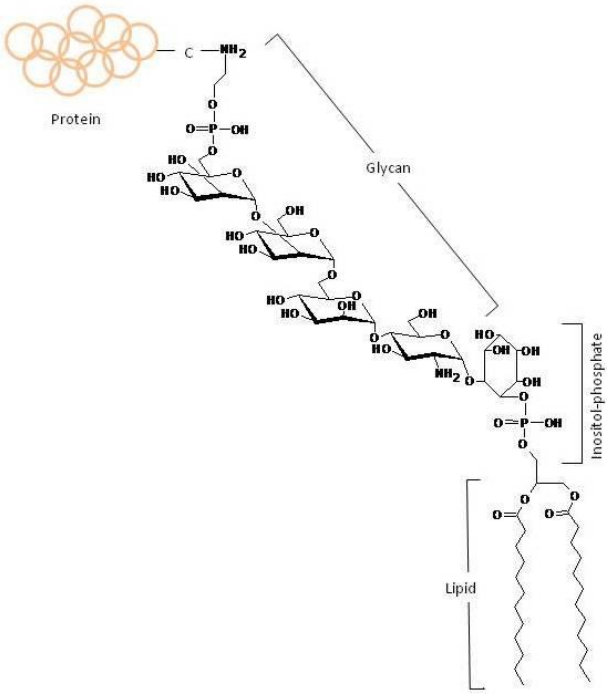


Figure 1.1 General structure of GPI anchors.

GPI linking is a post-translational modification (PTM) or the chemical modification of a protein after its translation. The PTM of proteins extends the range of functions of the protein by attaching to it other biochemical functional groups. The prediction and identification of PTMs in protein sequences is an important step in the annotation of proteomes since this can aid in developing the understanding of biological functions. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a powerful tool for identifying these PTMs. In high-throughput experiments, mass spectrometry (MS) database search engines, such as SEQUEST provide a ranked list of peptide identifications based on thousands of MS/MS spectra obtained in a MS experiment. However, since the GPIs are complex structures, these search results are not in themselves sufficient for confident assignment of GPI sites as identification of typical mass differences requires time-consuming manual assessment of the spectra by an experienced analyst. This examination requires interpretation of the fragmentation pattern in line with the experience of the scientist, and comparison of multiple interpretations of the data.

Proteins destined to be linked with GPIs contain a short amino acid sequence, called the omega-site (ω-site), that serve as a signal to attach the protein to the GPI. This feature has been used to develop computational methods based on Neural Networks (NNs) and Hidden Markov Models (HMMs) to predict the GPI-anchoring sequences with good accuracy (Fankhauser and Mäser, 2005). These prediction schemes are heavily dependent on training sets of experimentally confirmed protein sequences to be GPI-anchored. The main idea is to identify the ω-site of the protein and then classify it as a potential GPI linked protein. Omaetxebarria et al. (2007) used LC-MS/MS data for the identification of GPI-anchored proteins. They combined Bayesian networks with the use of the machine learning platform WEKA to make their predictions.

Recently, Nakayasu et al. (2009) identified 79 novel species of GPIs using manual interpretation of the fragmentation pattern with samples obtained from the parasite *T. cruzi* and analyzed by LC-MS/MS. Their method is the first large-scale analysis of the entire collection of free and protein-linked GPIs (i.e., the GPIome) of a eukaryote and resulted very effectively in characterizing complete GPI molecules. Their interpretation and decision process has a series of steps that can be captured in a computational model which in turn can speed up the analysis and obtain high-quality results.

The interpretation of mass spectrometry data is an integral part of structure elucidation of chemical compounds. In essence, there are two ways to look at the outcome of an LC-MS/MS experiment: examine the LC-MS/MS data as if it was a spectrum, showing the peak masses and intensities in a graph; or use the peak table, called DTA, generated by the LC-MS/MS software. The spectrum is a visual summary of the results of a mass spectrometry experiment (figure 1.2). The Y axis is labeled relative abundance. This is the abundance relative to the tallest peak in the spectra with the tallest peak set to 100%. The X axis is mass divided by charge, *m/z*. For example, if the mass of a molecule is 2000 Da and the molecule posses two proton adducts (double charge or MS^2) its *m/z* value is equal to (2000+2)/ 2, the *m/z* value read on the spectrum is 1001. This is the peak with heavier *m/z* value in the spectra also known as the "parent ion" and accounts for the entire *m/z* of a molecule. Peaks with smaller *m/z* than the

parent ion may represent the fragments of the molecule, or they may be just noise. The fragmentation pattern not only allows the determination of the mass of an unknown compound but also allows guessing the molecular structure. For some time, interpretation of the fragmentation patterns for the identification of GPI structures requires manual assessment of the spectra. In order to recognize GPI structures after an LC-MS/MS experiment, an experienced analyst performs the following series of steps:

1. Prefilter. Since all known GPI anchors must have at least three mannose residues, the analyst filters the obtained data from an LC-MS/MS experiment by the presence of fragments correspondent to this monosaccharide (shift of 162.052823 Da).
2. Look for familiar peaks. Once the data has been prefiltered, the analyst performs a visual inspection of the spectra and looks for the parent ion and peaks that resemble the mass and fragmentation pattern of a GPI molecule and selects various candidate spectra.
3. Chemical structure drawing. For each peak that resembles a part of a GPI, the analyst draws a chemical representation of this fragment using software such as ACD/ChemSketch. Adding up all the fragments the software calculates the mass of the structure which should be equivalent to the parent ion *m/z* value. This step is performed multiple times in order to take into consideration the possible variations of the molecular structure.
4. Data interpretation. After considering the different possibilities of a structure, the analyst decides the best candidate (or candidates) according to his experience and knowledge.

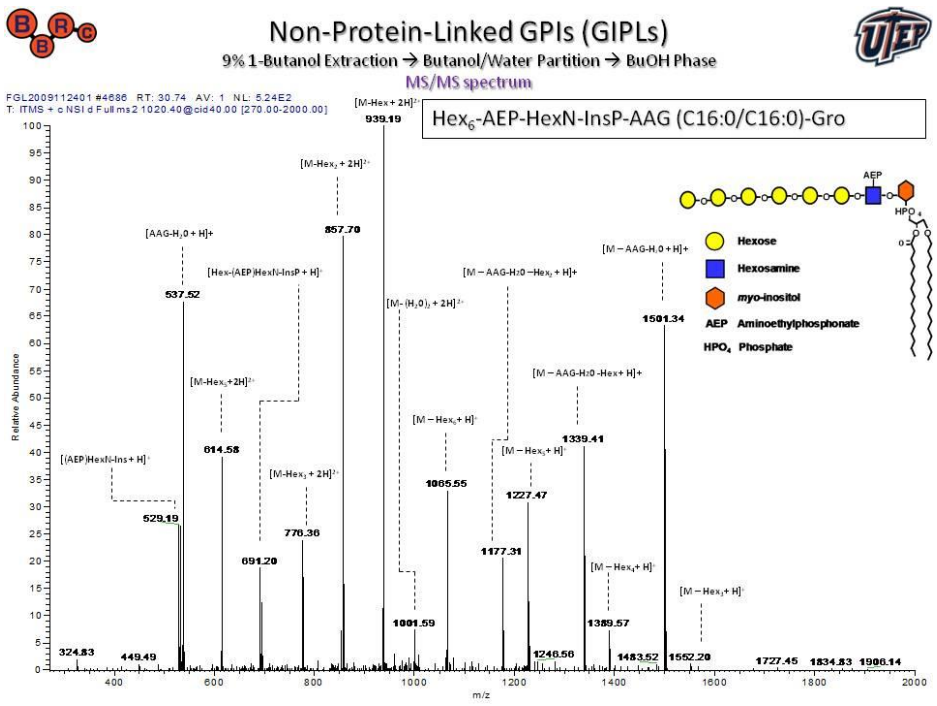


Figure 1.2 MS^2 spectra of GPI species Hex6-AEP-HexN-InsP-AAG(C16:0/C16:0)-Gro with annotation

A DTA file is a representation of the spectra that contains the parent mass, charge and observed fragmentation pattern in the form of peak lists. A DTA sample is provided in figure 1.3.

The aim of the current thesis is to provide a tool to automate the analysis and interpretation of the GPIomic MS data, attempting to address the following problems:

- **Problem 1.** How can we build a computational model to represent the analysis and decision process of an experienced scientist to substantially speed up and validate a large scale GPIomic experiment?
- **Problem 2.** Since there is no publicly available database of GPI structures, how can we construct one so we can search through it to identify the correct structure for a given spectrum?

- **Problem 3.** What are the analytical criteria that maximize the accuracy of the database search results? More specifically, we need to formulate a scoring system for predicted structures such that the correct structure is among the few highest scoring ones.

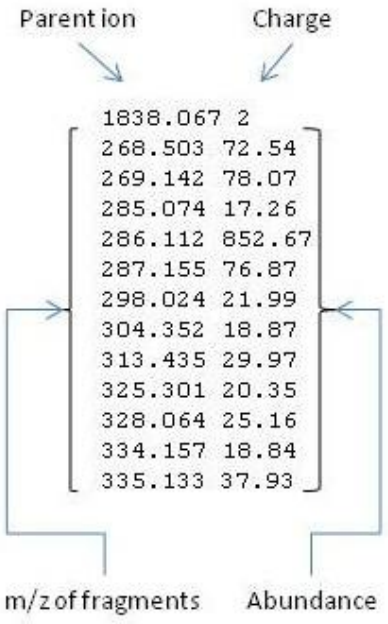


Figure 1.3 DTA sample generated by the LC-MS/MS software

The following summarizes my approach to these three problems.

Problem 1

In order to model the decision process of an analyst in determining the structure of a GPI molecule, I follow the same four steps as described above. The analysis strategy can begin with the reduction of the data (Step 1). If the mass and charge state of parent ions can be determined, it is possible to search for corresponding structures in a database of GPI structures (Steps 2 and 3). Fragments for each structure can be mapped against the DTA and then ranked to find the best match (Step 4).

The computational assessment of a DTA over the manual interpretation of a spectra should accelerate a GPIomic analysis both quantitatively and qualitatively; quantitatively because a large volume of LC-MS/MS data can be analyzed in less time; and qualitatively because the manual assessment of a spectrum can be influenced by subjective judgment, and researchers with less experience in this type of analysis can benefit with the aid of an automated process.

Problem 2

The biosynthesis of GPI molecules produces a core structure consisting of a lipid attached to a glycan as shown in Figure 1.1. The glycan group can be produced with a combination of a small number of monosaccharides along with inositol and phosphate. The lipid part of the molecule can have one of the four possible types of lipid tails: ceramide, lyso-acyl or lyso-alkyl glycerol (lyso), alkylacylglycerol (AAG) and diacylalkylglycerol (DAG). A data table with all the possible combinations of these two groups can be constructed, making it possible to scan a given DTA for a match between a parent ion and a structure. The construction of the glycan and lipid tables is explained in Chapter 3.

Problem 3

Our aim is to develop a tool that automates the processing of large numbers of spectra with high sensitivity and sufficiently good selectivity in the identification of GPI molecules. After each one of the candidate structures have gone through the process of finding a match in the DTA, they must be scored to determine the most probable structure (or structures). An explanation of the score system developed is presented in Chapter 4.

The thesis is organized as follows. In chapter 2, we introduce the biochemistry, LC-MS/MS analysis and review of the current computational prediction methods of GPIs. In chapter 3, we will discuss the LC-MS/MS data as the basic input for our analysis. We also explain the construction of the data tables. Chapter 4 will show our methodology for modeling large scale GPIomic experiments, including our algorithm and score system. Finally, in chapter 5 we present the results and discuss the effectiveness of the model.

Chapter 2: Literature Review

2.1 GPI anchor molecules

All living cells have a membrane that encloses their contents and serves as a semi-porous barrier to the outside environment. The cell membrane is composed of a bilayer of phospholipids which have a hydrophilic head and two hydrophobic tails. Within the phospholipid bilayer of the cell membrane, many different proteins are either peripheral proteins or integral membrane proteins. Some of these proteins have carbohydrates attached to their external surfaces and are, therefore, referred to as glycoproteins. A number of these glycoproteins contain also a lipid group, which help anchor the entire molecule to the cell membrane. A glycan part attached to a protein faces the extracellular environment (Figure 2.1). GPIs are present in eukaryotes and take part in significant biological processes such as cell-cell interactions and antigenic presentation (McConville and Ferguson, 1993).

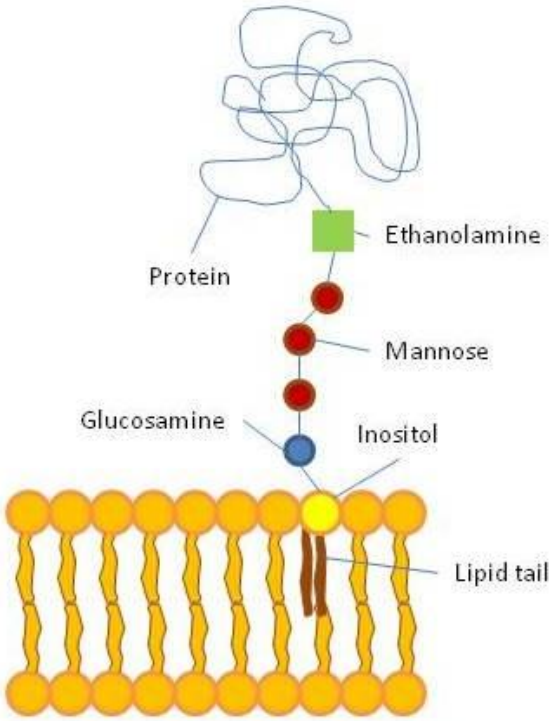


Figure 2.1 GPI anchored molecule attached to the cell membrane.

2.1.1 Biochemistry

The biosynthesis of GPIs requires a variety of enzymes, all of them acting in a defined and consecutive way. These reactions are carried out in the endoplasmic reticulum (ER). The final product posses a common structure consisting of a lipid tail attached to a glycan core (Figure 2.2). Modifications of this general structure make a number of variations of the molecule; these include extra mannose (Man) , ethanolaminephosphate (EtNP), and/or aminoethylphosphonate (AEP) residues substituting the glycan core, and/or an extra fatty acid (acyl) group attached to the myo-inositol ring, increasing the complexity of the GPI structure (Ferguson, 1999; McConville and Ferguson, 1993). Because of their complex structure and amphiphilic nature, GPIs are difficult to be extracted, purified, and fully characterized.

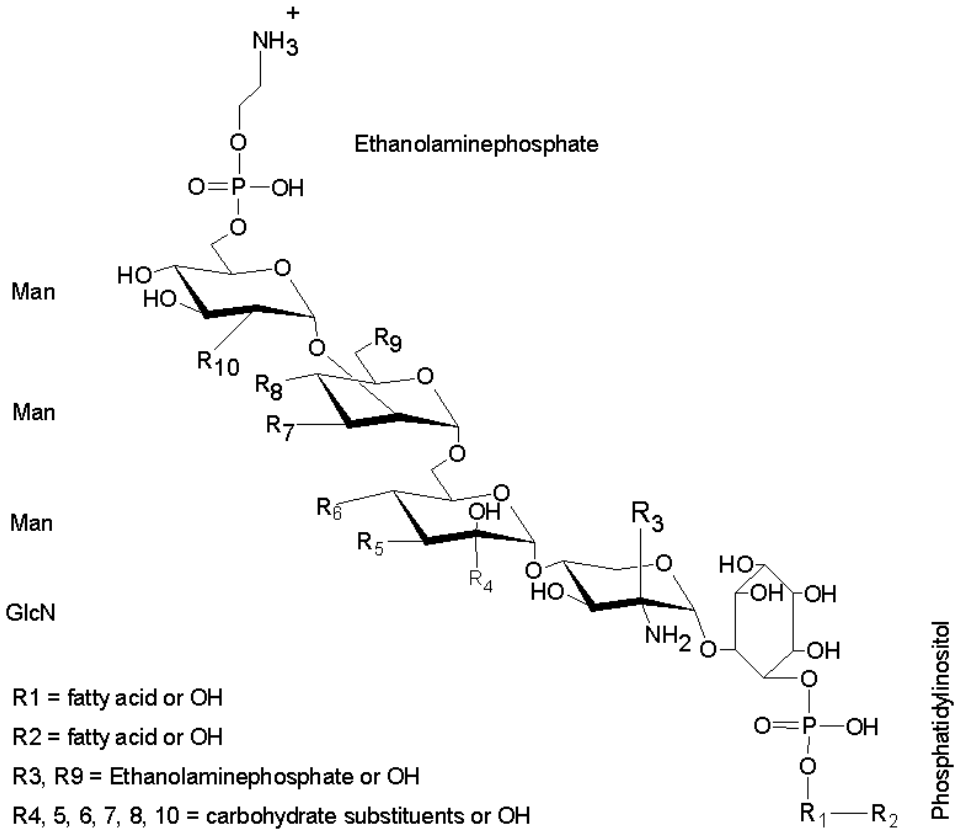


Figure 2.2 General structure of GPI anchors. Diversity of GPIs is derived from various substitutions of this core structure and is represented as R groups.

2.1.1.1 Glycan part

The term glycobiology joins the knowledge of carbohydrate chemistry and biochemistry with modern understanding of molecular biology of complex carbohydrates, which are often named glycans in this context. Glycans, therefore refer to the carbohydrate portion of a glycoconjugate, such as a glycoprotein, glycolipid or a proteoglycan.

In general, the glycan part of a GPI molecule consists of a phosphatidylinositol linked to a glucosamine (GlcN) residue followed by three Man residues. The terminal Man is connected to EtNP, which links in turn to the C-terminus of a protein.

The structural complexity of glycans, far greater than that of proteins and nucleic acids, allows them to encode information for specific molecular recognition and to determine protein folding, stability and pharmacokinetics (Morelle and Michalski, 2005). The glycan may be a single monosaccharide or an oligosaccharide. The attachment of glycans to a protein makes glycoproteins especially diverse. Glycans are linked to other biomolecules, such as lipids or amino acids, through glycosidic linkages to form glycoconjugates. In general, glycoproteins compose an assorted population of glycoconjugates containing between one and several dozen different glycans. The position and amount of glycans within a molecule makes it possible to form as many as 10^{12} distinct structures from as few as six different monosaccharide units (Morelle and Michalski, 2005). Therefore, in order to elucidate the structure of a particular glycan it is necessary to determine the sequence, composition and branching of its monosaccharide units as well as its glycosidic linkages and anomeric configuration. For that reason, the challenges of analytical glycobiology are much greater than those encountered in genomics and proteomics.

2.1.1.2 Lipid part

Lipids are a diverse class of biological molecules that play a central role as structural components of biological membranes, energy reserves, and signaling molecules (Yetukuri, et al., 2007). Lipids are structurally highly diverse because of the many possible variations of the

lipid building blocks, how these blocks are linked and their variation of both chain length and degree of saturation. Yetukuri, et al. (2007) estimated that the theoretical number of lipids covering major lipid classes is close to 200,000.

There are four classes of lipid tails in a GPI molecule: ceramide, lyso-acyl or lyso-alkyl glycerol (lyso), alkylacylglycerol (AAG) and diacylalkylglycerol (DAG). In figure 2.3 we show some lipid tails that can be present in GPIs.

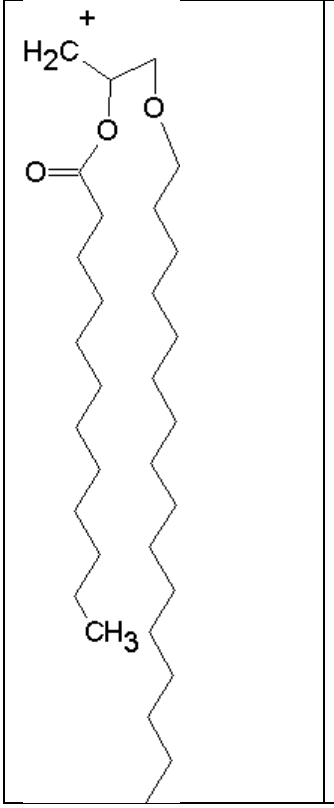
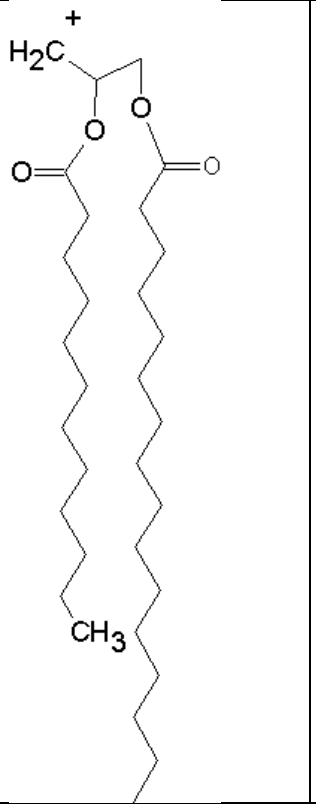
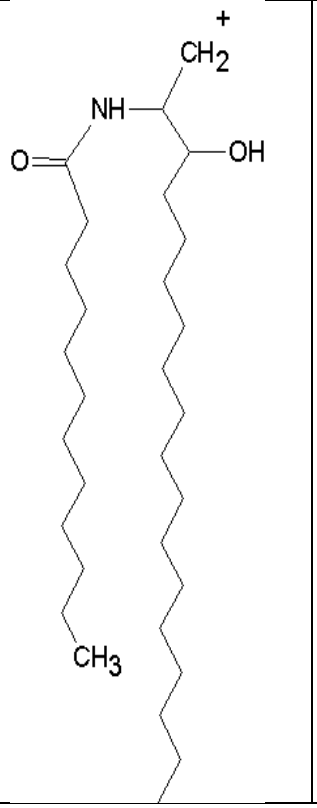
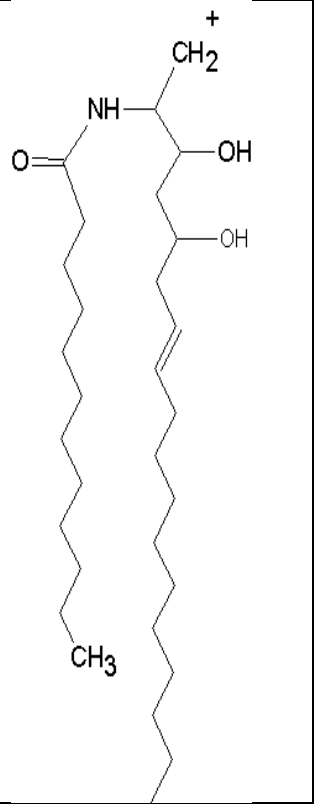
AAG-C16:0/C12:0	DAG-C16:0/C12:0	Ceramide-C12:0/d18:0	Ceramide-C12:0/t18:1
			

Figure 2.3 Possible arrangements of lipid tails in GPIs

2.1.1.3 GPI diversity

Isobaric ions are ions that have the same nominal mass but different exact mass. For example N_2, C_2H_4 and CO all have a nominal mass of 28 Da. Their exact masses are: N_2 = 28.00615 Da, C_2H_4 = 28.0313 Da and CO = 27.99491. Because of the variety of the structures, modification and branching GPIs are often isobaric (Figure 2.4).

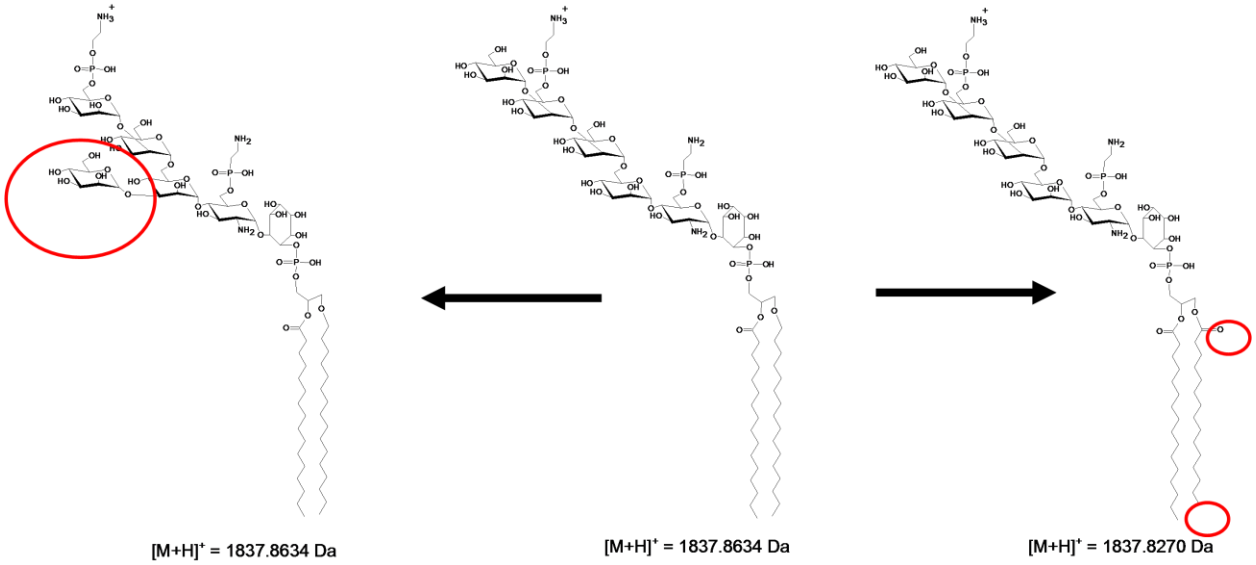


Figure 2.4 Isobaric GPIs

2.1.2 LC-MS/MS analysis

The study of glycoconjugates is a challenging task because these molecules exhibit complex structures that can differ in linkage and the level of branching. MS is one of the most powerful and versatile techniques for the structural analysis of glycoconjugates. The characterization of glycoproteins by mass spectrometry is typically more difficult than the mass spectrometric analysis of proteins, because glycoproteins exhibit extensive heterogeneity and because they are ionized less efficiently than proteins (Morelle and Michalski, 2005).

LC-MS^n has been successfully used for the characterization of GPI molecules (Nakayasu et al., 2009). The structure of GPIs can be elucidated using MS, exposing the types of glycans present

and providing evidence of structures that are potentially important for biological function. Monosaccharide sequences, branching, and linkages can be determined through fragmentation. The observed fragments depend on factors such as the type of ion ([M+H]+, [M+Na]+ etc.), its charge state, and the time available for the fragmentation (retention time). In general, glycans fragment to give two major types of ions. These ions are the result of two types of cleavage: glycosidic cleavages where bond rupture occurs between the sugar rings and involves a hydrogen migration and cross-ring cleavages that involve the rupture of two bonds on the same monosaccharide. The nomenclature generally used for describing these fragment ions is that proposed by Domon and Costello (1988) (Fig. 2.5). Glycosidic cleavages provide information on constituent monosaccharide sequence and branching. Crossring cleavages are usually weaker. The fragmentation of glycans through LC-MS^n is usually performed by MS^2.

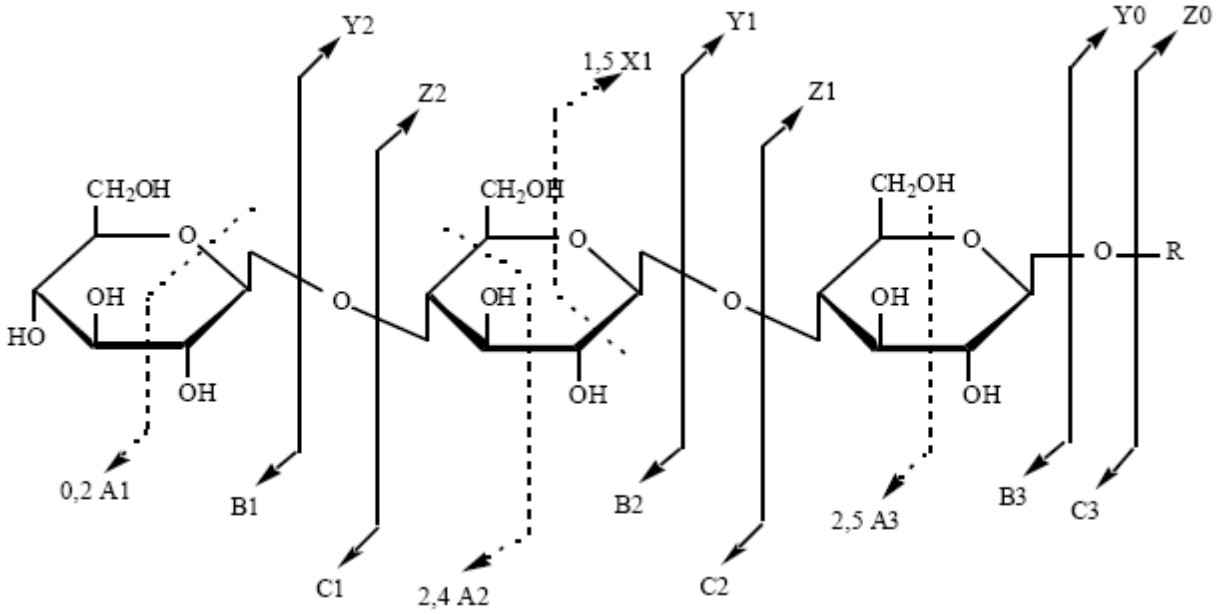


Figure 2.5 Nomenclature for describing the fragmentation of carbohydrates, after Domon and Costello (1988)

Although the lipids commonly produce specific fragmentation patterns by tandem MS, such information is not always available for each specie analyzed by lipidomics experiments (Niemela et al., 2009). Due to the structural characteristics of lipids their identification from fragment

mass spectra requires the use of MS^3. We, however, receive data from MS^2 experiments in which the fragments of glycans can be present but the lipids are not fragmented, they only appear as single peaks. Then we are aiming to interpret GPIs with a fragmented glycan and the entirety of the lipid part.

2.2 Computational Prediction

Bioinformatics resources for glycomics and lipidomics are very poor as compared with those for genomics and proteomics which are widely available for molecular biologists. Tools for the characterization of glycans and lipids have been built separately, but essentially they mirror the way in which proteomic characterization has been performed. There are two main classes of algorithms for the identification of all these compounds given LC-MS/MS data: database search algorithms and de novo sequencing algorithms. Here we will describe both.

2.2.1 Database search algorithms

In proteomics, various algorithms and computer programs have been developed for the identification of protein sequences using MS data to search over a database of known proteins. Corresponding *m/z* values within the database are scored and ranked to find the best match between a protein and the spectra. MS data are submitted to these programs in the form of peak lists (DTAs). One of the first programs of this kind which is still widely used is SEQUEST (Eng et al., 1994). In SEQUEST the interpretation of a spectrum for an unknown protein sequence proceeds by identifying a consecutive series of fragment ions whose differences correspond to residue masses for amino acids. In general there are four steps in the identification of proteins using this software. On the basis of mass alone, SEQUEST searches a database for candidate peptides. A virtual spectrum is created for each of these peptides and then it checks whether it matches the observed spectrum. The final outcome is a list with the best peptides that matched the spectra; each peptide is shown with their corresponding score.

MASCOT (Perkins et al., 1999) is another database search algorithm. MASCOT's fundamental approach is to calculate the probability that the observed match between the experimental

data set and each sequence database entry is a chance event. The match with the lowest probability is reported as the best match. Intensity information is ignored because peak intensities depend on the physical and chemical properties of the samples.

Since glycan databases are very small when compared to those existing in proteins and genomes, theoretical databases have been created to mimic searching algorithms used in proteomics. For example, Joshi et al. (2004) generated a theoretical database of glycan fragments. Using the 1674 fully characterized carbohydrate structures from GlycoSuiteDB (Cooper et al., 2003) they systematically produced a database with 3×10^6 fragments taking into consideration the Domon and Costello notation. With this database a searching algorithm finds the set of all candidate structures that have a fragment mass within a tolerance for each observed mass. Then, the union of the sets of candidate structures accumulated is found counting the number of times each carbohydrate structure is observed. Finally the structures are sorted by the number of times each carbohydrate structure was observed in order of most number of hits to least.

Another method for annotating possible N-glycan structures synthesized by mammals is to generate libraries based on biosynthetic rules and patterns, also called cartoons (Goldberg, et al., 2005). A table of potential glycans is used to match the peaks with cartoons within a tolerance. The cartoons provide compositional information and predictions of probable topologies. When the set of cartoons is obtained, a confidence score is assigned to each cartoon, hence the number of matching structures is greatly reduced

Lipid identification strategies also include the generation of databases with theoretically possible lipids, with information on masses, isotope patterns and additional constraints such as retention time (Yetukuri et al., 2007).

2.2.2 De novo sequencing algorithms

The de novo peptide sequencing problem is the reconstruction of a peptide sequence from LC-MS/MS data without the aid of a database from which known peptides can be matched. In the

process of collision-induced dissociation (CID), a peptide bond at a random position is broken into complementary ions, typically N-terminal ions called b-ions (prefix) and C-terminal ions called y-ions (suffix). Intermediate bonds are also broken (Figure 2.6). These ions are complementary because joining them determines the original peptide sequence.

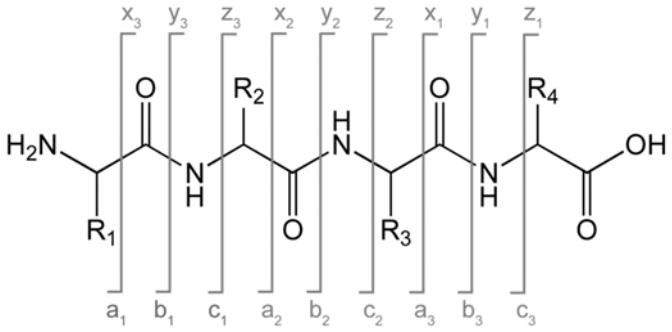


Figure 2.6 Nomenclature for describing the fragmentation of peptides

The interpretation of a spectrum basically deals with two factors (Chen et al. 2001): (1) it is unknown whether a mass peak corresponds to a prefix or a suffix subsequence; (2) some ions may be lost in the experiments and the corresponding peaks disappear in the spectra. In practice noise and other factors can affect a spectra. An ion may display two or three different peaks because of the distribution of isotopic carbons in the molecules (i.e., C^{12} and C^{13}). An ion may lose a water molecule or an ammonia molecule and display a different peak from its normal one. Also, the fragmentation may result in some other ion types such as a- and z-ions.

Dancik et al. (1999) approached the de novo sequencing problem by the following method. First, the spectral data is transformed to a directed acyclic graph, called a spectrum graph, where a node corresponds to a mass peak; an edge connects two nodes that differ by the total mass of possible amino acids in the nodes. Each node represents a possible prefix subsequence for the peak. Then, an algorithm is called to find the highest-scoring path in the graph or all paths with scores higher than some threshold. The concatenation of edge labels in a path gives

one or multiple candidate peptide sequences. The longest path in the resulting directed acyclic graph is the path that best explains the spectra.

Glyco-Peakfinder (Maass et al., 2007) is a web-service for the de novo determination of the composition of glycans. The algorithm requires a set of mandatory masses including the possible occurring monosaccharides, the cross-ring fragments for each monosaccharide, modifications that can occur at the reducing end, and charged ions. The calculation basically is a consecutive addition of mass increments of monosaccharides and losses of other small molecules that lead to core structure and depends on user settings. The algorithm provides all possible results for each given m/z value from a mass list independently from other peaks.

Chapter 3: Data

3.1 LC-MS/MS data

The mass spectrum consists of a collection of *m/z* values. The largest peak in the mass spectrum (100% relative intensity) is called the base peak. Graphically, *m/z* values are plotted along the horizontal axis and relative peak intensities along the vertical axis (Figure 1.2). The same data may be presented in tabular form (DTA), such as that in Figure 1.3. A DTA is a text file that contains two columns. In the first row the parent ion is at the left column and the charge at the right column. The rest of the rows show the *m/z* of the fragments in the first column and the abundance in the second column. Even though the graphical presentation is more visually instructive and its use is common in the analysis by visual inspection, the tabular presentation often contains additional information that is not readily apparent from the graphic display and is a high-quality input for a computer program.

3.1.1 Ions vs. Peaks

In MS it is important to distinguish between the terms ions and peaks. Ions are particles that have both mass and charge, and they can fragment to form other ions. There can be large or small numbers of ions, so that it is appropriate to speak of their relative abundance. On the other hand, peaks in a mass spectrum correspond to localized maximum signals produced by the detector and have only m/z values associated with them. These signals are either weak or strong (depending on the numbers of ions produced) and therefore are best described as having intensity. The abundance of peaks implies that there are many peaks, not that a given peak is big or little.

3.2 Data tables

In chapter 2 we described the intrinsic difficulties in elucidating the structure of glycolipids. Because the biosynthesis of glycolipids is not under direct genetic control, different enzymes are involved in the synthesis of sugar chains attached to proteins or lipids. This can lead the cell

in the production of several different glycan and lipid chains, many of them unknown. The availability of a comprehensive GPI database is a prerequisite for successfully developing a computational tool aimed at deciphering new, so far unknown GPI structures. When compared to genomics and proteomics, glycan databases are very poor, in particular, experimentally annotated GPIs account for only a few hundreds. Therefore, we developed our own data table of theoretical GPIs. In the next section we describe how we built our data table.

3.2.1 Glycan table

The glycan group can be produced with a combination of a small number of monosaccharides along with inositol and phosphate. Then, a small data table with the constant mass values of each of these compounds was created.

Table 3.1 Table with constant mass values for glycan fragments, inositol and phosphate

Fragment	Mass (Da.)
Hex	162.052823
dHex	146.057909
EtNP	123.008529
AEP	107.013614
HexN	161.068808
HexNAc	203.079373
Ins	162.052823
P	97.976895
NANA	291.095417

The most elementary glycan part of a GPI can be composed of three Hex residues, and one HexN residue followed by inositol-phosphate. This basic structure and its mass served as starting point to add all possible monosaccharides and construct a glycan table. The basic table consists of 72 rows and 40 columns (Table 3.2 shows only part of it).

Table 3.2 Sample of basic glycan table

	Hex(n)-HexN-Ins-P	EtNP-Hex(n)-HexN-Ins-P	Hex(n)-AEP-HexN-Ins-P
Hex3	907.256996	1030.265525	1014.270611
Hex4	1069.309796	1192.318325	1176.323411
Hex5	1231.362596	1354.371125	1338.376211
Hex6	1393.415396	1516.423925	1500.429011
Hex7	1555.468196	1678.476725	1662.481811
Hex8	1717.520996	1840.529525	1824.534611
dHex-Hex3	1053.314896	1176.323425	1160.328511

We combined columns and rows of the basic table to obtain structures and mass values of 2880 glycans (Table 3.3).

Table 3.3 Portion of the glycan table

Glycan + Inositol-phosphate	Mass (Da.)
Hex3-HexN-Ins-P-	907.257
Hex4-HexN-Ins-P-	1069.31
Hex5-HexN-Ins-P-	1231.363
Hex6-HexN-Ins-P-	1393.415
Hex7-HexN-Ins-P-	1555.468
Hex8-HexN-Ins-P-	1717.521
dHex-Hex3-HexN-Ins-P-	1053.315
dHex-Hex4-HexN-Ins-P-	1215.368
dHex-Hex5-HexN-Ins-P-	1377.42

3.2.2 Lipid tables

For each one of the four kinds of lipids of a GPI, the most basic structure and its mass was obtained. Then, adding all the possible carbons and unsaturations we acquired the structure and mass for each possible lipid (Table 3.4).

Table 3.4 Portion of lipid tables

Lipid	Mass (Da.)
AAG-28:0	498.4648
AAG-28:1	496.4492
AAG-28:2	494.4335
AAG-28:3	492.4179
AAG-28:4	490.4022

Lipid	Mass (Da.)
DAG-28:0	512.4441
DAG-28:1	510.4284
DAG-28:2	508.4128
DAG-28:3	506.3971
DAG-28:4	504.3815

Lipid	Mass (Da.)
12:0/d18:0	483.47
13:0/d18:0	497.48
14:0/d18:0	511.5
15:0/d18:0	525.51
16:0/d18:0	539.53

Lipid	Mass (Da.)
12:0/lysoacylglycerol	247.21
13:0/lysoacylglycerol	261.23
14:0/lysoacylglycerol	248.21
18:1/lysoakylglycerol	342.31
19:1/lysoakylglycerol	356.33

Our final table is the combination of the glycan and lipid tables and in total there are 2,592,000 structures. Table 3.5 shows a sample of the final table including the lowest and highest mass values.

Table 3.5 Sample of table including theoretical GPI structures and its masses

Structure	**Mass Value (Da)**
Hex3-HexN-Ins-P-12:4/lysoakylglycerol	1142.426
Hex3-HexN-Ins-P-12:3/lysoakylglycerol	1144.442
Hex3-HexN-Ins-P-12:2/lysoakylglycerol	1146.458
Hex6-HexN-Ins-(C18:0)-P-18:3/t18:1	2056.117
Hex6-HexN-Ins-(C18:0)-P-18:4/t18:0	2056.117
Hex6-HexN-Ins-(C18:1)-P-18:1/t18:2	2056.117
EtNP3-Hex4-HexN-Ins-P-AAG-34:1	1839.823
EtNP3-dHex-Hex3-HexN-Ins-P-DAG-34:0	1839.823
HexNAc-dHex2-Hex4-HexN-Ins-P-18:3/t18:2	2120.982
HexNAc-dHex2-Hex4-HexN-Ins-P-18:4/t18:1	2120.982
EtNP-NANA-HexNAc-dHex2-Hex8-AEP-HexN-Ins-(C18:0)-P-DAG-52:1	3845.891
EtNP-NANA-HexNAc-dHex2-Hex8-AEP-HexN-Ins-(C18:1)-P-DAG-52:0	3845.891
EtNP-NANA-HexNAc-dHex2-Hex8-AEP-HexN-Ins-(C18:0)-P-DAG-52:0	3847.907

Chapter 4: Methods

4.1 Modeling Large Scale GPIomics Experiments

Because the biosynthesis of GPIs requires a variety of enzymes, all of them acting in a defined and consecutive way, there are currently no methods available to amplify these molecules similar to DNA amplification using polymerase chain reaction (PCR) techniques. Consequently, highly sensitive analytical chemistry methods have to be applied. This is a difficult task and few GPIs have been completely characterized, although large-scale screening of samples is a realistic possibility (Nakayasu et al., 2009).

Many GPIs are very similar, or are isometric forms, and therefore difficult to separate and may be present in only low amounts. The peaks separated in LC-MS/MS profiles cannot generally be identified directly as several GPIs may have the same properties. For example, they might have the same mass. Hence, the identification of GPI structures requires detailed knowledge of the properties of the molecule and its fragmentation patterns. So far, this process has been performed manually by experienced analysts.

GPIomic analysis is comparable to proteomic analysis. However the development of an algorithm will differ due to the possible presence of branch points, linkages and anomericity in GPI molecules. In peptide sequencing, the fragmentation is across the peptide bond and a linear sequence can be matched to a database of known proteins. Similarly, in order to identify GPI structures, LC-MS/MS data can be matched against a theoretical database containing a set of all theoretically possible GPIs. The key to this methodology, as in protein sequencing, is the criteria established for ranking the potential GPI structures that are compatible with the observed spectra.

4.2 Algorithm

In this research we present an algorithm for the identification of GPI structures using LC-MS/MS data. In general, the algorithm consists of five steps. The analysis strategy begins with the reduction of the data (Step 1). The mass and charge state of parent ions are determined (Step 2) in order to search for corresponding structures in the theoretical data table (Step 3). Fragments for each structure are mapped against the DTA (Step 4). The final step is the ranking of structures (Step 5). The steps are explained below in more detail and an example is shown in Figure 4.1.

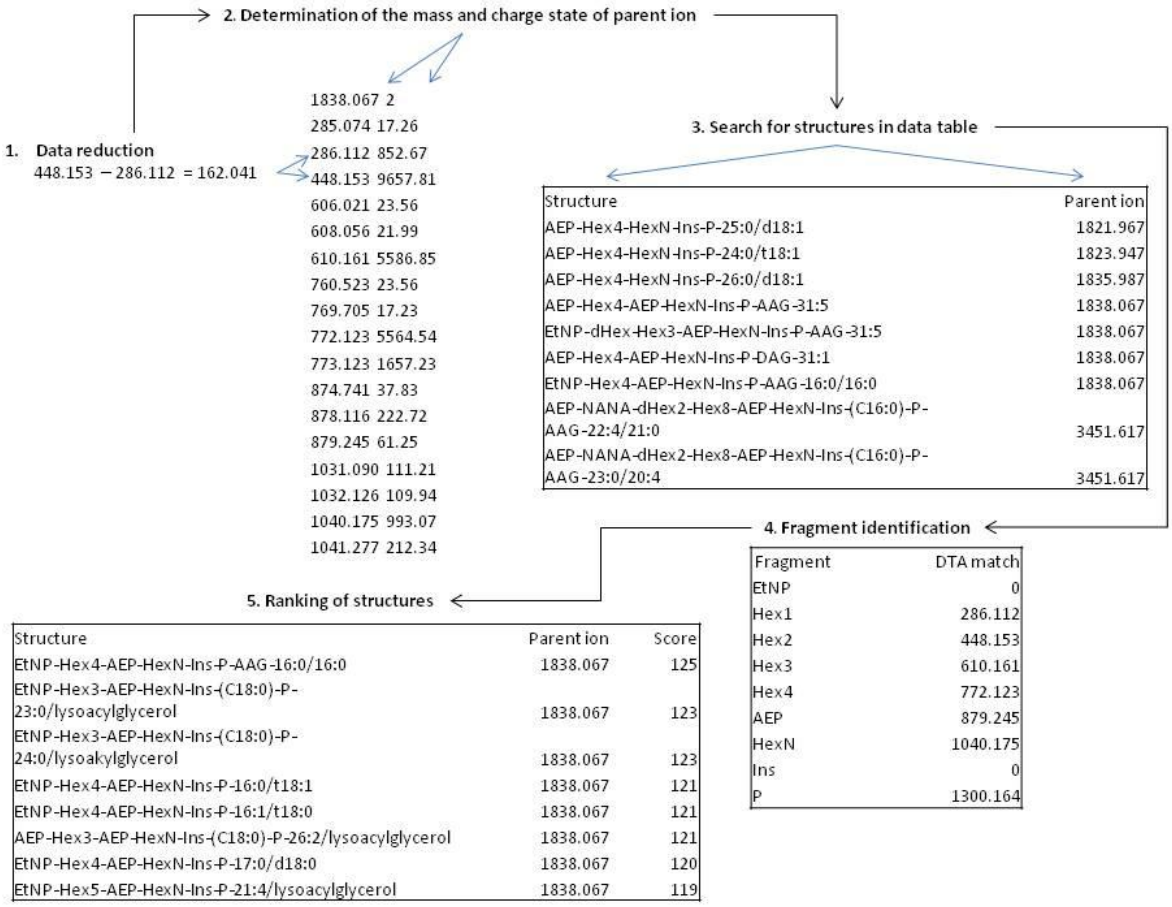


Figure 4.1 example of the steps required to obtain a GPI structure

4.2.1 Data reduction. High-throughput profiling of GPIs generates very large amounts of data. All these data, however, can be used meaningfully only if we can extract the information pertaining to the fragmentation patterns of GPI molecules. The presence of mannose fragments is an indication that the data can be used to characterize entire GPIs. The algorithm first identifies DTAs by the presence of fragments corresponding to this monosaccharide (shift of 162.052823 Da).

4.2.2 Determination of the mass and charge state of parent ions. For each DTA, the algorithm finds the parent ion and charge. This information reveals the whole electrically charged molecule that dissociates to form the fragments whose information occupies the remaining rows of the file.

4.2.3 Search for structures. To match a spectrum to a structure in the data table, the *m/z* of the parent ion is scanned throughout the data table to find all possible structures with a tolerance of ±2Da of the *m/z* value. The *m/z* value is a quantity formed by dividing the mass number of an ion by its charge number, for example, for the ion $C_7H_7^{+2}$, *m/z* equals 45.5. The *m/z* calculation also provides information of the possible fragmentation pattern of the parent ion ($[F+H]^+$, $[F+Na]^+$, $[F+2H]^{+2}$, etc).

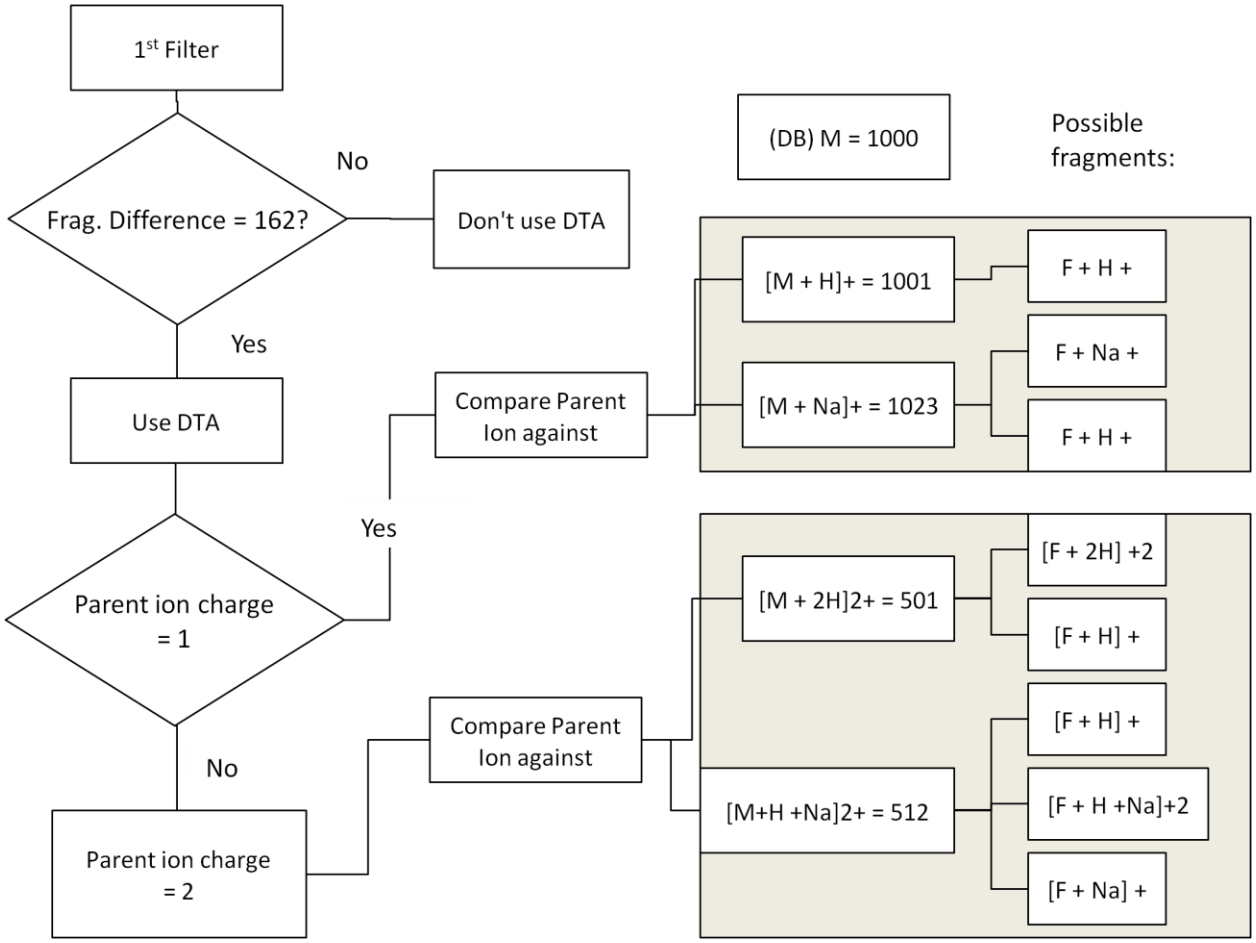


Figure 4.2 Flowchart of steps 1, 2 and 3 with mass example

4.2.4 Fragment identification. For each candidate structure the algorithm iterates through the fragments to find an occurrence within the DTA with a tolerance of ±1Da of the value. A lookup table of theoretical fragments is used to acquire the mass value of the current fragment (Table 1.1 and 1.3).

4.2.5 Score system. The final and most important step in the algorithm is to rank the possible GPI structures. At present each candidate structure is evaluated by counting the number of fragment matches within the DTA. The iteration process described above is performed in a forward and backward fashion. In the forward prediction the accumulative masses of each fragment are evaluated and for each match found the score is increased. Similarly, the backward prediction attempts to "remember" what a past value was and for each match the

score is also increased. In summary, the number of predicted fragments that match fragments observed in the DTA within ±1Da are summed.

Diagnostic fragment ions are also considered. A diagnostic ion is a fragment ion that is sufficient to identify a molecular species. Fragment-ions corresponding to the neutral loss of Hex residue(s) and AEP or EtNP are highly abundant in *T. cruzi*, for instance, at *m/z* 529.3 one of the most abundant ions is the fragment corresponding to the *inositolphosphate* (InsP) attached to AEP-HexN (AEP-HexN-InsP) (Nakayasu et al., 2009). EtNP linked to the first mannose residue of the molecule is another major feature of the GPI structure (Vainauskas and Menon, 2006). If a diagnostic ion is present in the structure, the score is incremented by 100 points. The neutral losses of water are also considered and if a fragment $-2H_2O$ is found, the score is incremented by one point.

A Perl script was written to implement the algorithm. The computer used to run the script was a Fujitsu server with two quad-core Xeon (8 cores total) CPU, 8GB RAM memory and RedHat Enterprise Linux 5.

Chapter 5: Results, Discussion, and Future Work

5.1 Results

The overall effectiveness of our algorithm was tested with a set of 79 GPI structures that had been previously manually annotated (Nakayasu et al., 2009). We also used the set of spectra from the same experiments.

Table 5.1 illustrates a results report for one predicted structure. The details of this report are as follows:

- At the top the parent-ion mass for the structure is indicated, the structure itself and the possible fragmentation pattern of the parent ion.
- Next, two tables specify the forward and backward compositions of the structure whose theoretical masses (predicted) match the observed masses found in the DTA.
- Each value is reported with the corresponding cumulative mass of the fragment including its loss of water.
- At the end an indication if a structure contains a diagnostic ion, and the final score.

Table 5.1 Complete results for one structure

EtNP-Hex4-AEP-HexN-Ins-P-AAG-16:0/16:0 1838.067 Charge = 2 [M + H]+
AEP-HexN-InsP found in prediction Score: 129

Backward Prediction:	Predicted	Observed	Pred M-H2O	Obs M-H2O
AAG-16:0/16:0	537.526	537.479	519.516	510.492
HexN	958.625	958.392	940.614	0
AEP	1065.638	1065.458	1047.628	1048.327
Hex1	1227.691	1227.366	1209.681	1209.542
Hex2	1389.749	1389.4242	1371.739	1371.6
Hex3	1551.807	1551.4825	1533.798	1533.658
Hex4	1713.866	1713.5407	1695.856	1695.717
EtNP	1836.874	1836.5492	1818.864	1818.725

Forward prediction:	Predicted	Observed	Pred M-H2O	Obs M-H2O
EtNP	0	0	0	0
Hex1	286.112	852.6	268.503	72.5
Hex2	448.153	9657.8	430.205	348.2
Hex3	610.161	5586.8	593.152	50.2
Hex4	772.123	5564.5	755.084	160
AEP	879.245	61.2	861.32	28.3
HexN	1040.175	993	1022.256	22
Ins	0	0	0	0
P	1300.164	13559.9	1283.125	233.7

A simpler table is generated separately to summarize the information for various structures predicted and indicates the name of the file, the structure, parent-ion *m/z* and score (Table 5.2).

Table 5.2 Tabular list of predicted structures

Structure	Parent-ion	Score
EtNP-Hex4-AEP-HexN-Ins-P-AAG-16:0/16:0	1838.067	129
EtNP-Hex3-AEP-HexN-Ins-(C18:0)-P-23:0/lysoacylglycerol	1838.067	123
EtNP-Hex3-AEP-HexN-Ins-(C18:0)-P-24:0/lysoakylglycerol	1838.067	123
EtNP-Hex4-AEP-HexN-Ins-P-16:0/t18:1	1838.067	121
EtNP-Hex4-AEP-HexN-Ins-P-16:1/t18:0	1838.067	121
AEP-Hex3-AEP-HexN-Ins-(C18:0)-P-26:2/lysoacylglycerol	1838.067	121
EtNP-Hex4-AEP-HexN-Ins-P-17:0/d18:0	1838.067	120
EtNP-Hex5-AEP-HexN-Ins-P-21:4/lysoacylglycerol	1838.067	119

We were able to correctly match MS/MS spectra with their proposed GPI structure using our testing set of structures. We evaluated the effect on search accuracy by using the candidate structures among the top 5, 10, 15 and 20 structures ranked with the highest scores. Selecting the candidates with the top 15 scores resulted in the identification of all except 5 of the 78 testing structures. However, this means that for every correctly predicted structure, there are at least 14 false positives. The main problem is that our scoring method produces an overlap between the scores of correct and incorrect structures identified since a number of isobaric structures exist within the database.

Table 5.3 Experimentally confirmed GPI structures (Nakayasu et al., 2009) and the rankings and scores of the correctly predicted structures.

Species	Observed structure	Parent ion m/z value	Predicted structure	Score	Rank
1	[EtNP]Hex5-[AEP]HexN-InsP-C34:0(OH)3-Cer	2002.9	EtNP-Hex5-AEP-HexN-Ins-P-16:0/d18:1	104	2
2	[AEP]Hex6-HexN-InsP-C34:0(OH)3-Cer	2041.96	Hex6-AEP-HexN-Ins-P-16:4/t18:2	19	8
3	[EtNP]Hex5-[AEP]HexN-InsP-C16:0/d18:1-Cer	1984.9	EtNP-Hex5-AEP-HexN-Ins-P-16:0/d18:1	104	2
4	[AEP]Hex8-HexN-InsP-C16:0/d18:1-Cer	2348.06	AEP-Hex8-AEP-HexN-Ins-P-16:0/d18:2	101	5
5	[AEP]Hex5-[AEP]HexN-InsP-C16:0/d18:1-Cer	1968.9	AEP-Hex5-AEP-HexN-Ins-P-16:0/d18:0	122	2
6	Hex6-[AEP]HexN-InsP-C16:0/d18:1-Cer or [AEP]Hex6-HexN-InsP-C16:0/d18:1-Cer	2023.94	Hex6-AEP-HexN-Ins-P-16:0/d18:0	104	4
7	[AEP]Hex7-HexN-InsP-C34:1(OH)2-Cer	2186	Hex7-AEP-HexN-Ins-P-15:0/d18:2	107	7
8	[AEP]Hex6-[AEP]HexN-InsP-C34:1(OH)2-Cer	2130.96	AEP-Hex6-AEP-HexN-Ins-P-16:0/d18:0	102	4
9	[AEP]Hex4-HexN-InsP-C16:0/d18:1-Cer	1699.84	Hex4-AEP-HexN-Ins-P-16:0/d18:1	116	4
10	[AEP]Hex8-HexN-InsP-C16:0/d18:1-Cer	2372.06	AEP-Hex8-AEP-HexN-Ins-P-16:0/d18:2	101	5
11	[EtNP]Hex5-[AEP]HexN-InsP-C16:0/d18:0-Cer	1986.92	EtNP-Hex5-AEP-HexN-Ins-P-16:0/d18:1	104	2
12	[AEP]Hex7-HexN-InsP-C34:0(OH)2-Cer	2188	Hex7-AEP-HexN-Ins-P-14:2/t18:1	108	6
13	[NANA]Hex7-[AEP]HexN-InsP-C34:0(OH)2-Cer	2499.1	NANA-Hex7-AEP-HexN-Ins-P-16:4/t18:2	115	18
14	[EtNP]Hex6-[AEP]HexN-InsP-C34:0(OH)2-Cer	2148.98	EtNP-Hex6-AEP-HexN-Ins-P-14:0/t18:2	108	3
15	[AEP]Hex6-HexN-InsP-C34:0(OH)2-Cer or Hex6-[AEP]HexN-InsP-C34:0(OH)2-Cer	2025.96	Hex6-AEP-HexN-Ins-P-16:0/d18:0	104	4
16	[AEP]Hex5-[AEP]HexN-InsP-C16:0/d18:0-Cer	1970.92	AEP-Hex5-AEP-HexN-Ins-P-16:0/d18:0	121	2
17	[AEP]Hex4-HexN-InsP-C34:0(OH)2-Cer	1701.86	Hex4-AEP-HexN-Ins-P-16:1/d18:0	116	4
18	[AEP]Hex6-HexN-InsP-alkyl-C16:0-acyl-C16:0-Gro	2040.96	Hex6-AEP-HexN-Ins-P-AAG-31:0	101	4
19	[EtNP]Hex5-[AEP]HexN-InsP-alkylacyl-C32:0-Gro	2001.92	EtNP-Hex5-AEP-HexN-Ins-P-AAG-32:0	104	3
20	[AEP]Hex8-HexN-InsP-alkylacyl-C32:0-Gro	2365.06	Hex8-AEP-HexN-Ins-P-AAG-32:0	100	4
21	[AEP]Hex5-[AEP]HexN-InsP-alkylacyl-C32:0-Gro	1985.92	Hex5-AEP-HexN-Ins-P-AAG-33:0	104	8
22	[EtNP]Hex4-[AEP]HexN-InsP-alkylacyl-C32:0-Gro	1839.86	EtNP-Hex4-AEP-HexN-Ins-P-AAG-32:0	114	8
23	[AEP]Hex4-HexN-InsP-alkylacyl-C32:0-Gro	1716.86	Hex4-AEP-HexN-Ins-P-AAG-32:0	101	3
24	[AEP]Hex4-HexN-InsP-alkylacyl-C34:2-Gro	1740.86	Hex4-AEP-HexN-Ins-P-AAG-34:2	103	1
25	[AEP]Hex6-HexN-InsP-alkylacyl-C34:1-Gro	2066.98	Hex6-AEP-HexN-Ins-P-AAG-35:5	101	3
26	[AEP]Hex4-HexN-InsP-alkylacyl-C34:1-Gro	1764.84	Hex4-AEP-HexN-Ins-P-AAG-34:0	114	16
27	[AEP]Hex6-HexN-InsP-alkylacyl-C34:0-Gro	2067.98	Hex6-AEP-HexN-Ins-P-AAG-35:5	101	3

28	[AEP]Hex6-HexN-InsP-C22:0/d18:1-Cer	2107.02	Hex6-AEP-HexN-Ins-P-21:4/d18:1	101	3
29	[EtNP]Hex5-[AEP]HexN-InsP-C22:0/d18:1-Cer	2068.98	EtNP-Hex5-AEP-HexN-Ins-P-21:4/d18:2	104	6
30	[EtNP]Hex5-[AEP]HexN-InsP-C40:0-(OH)2-Cer	2071.02	EtNP-Hex5-AEP-HexN-Ins-P-22:4/t18:2	104	2
31	[EtNP]Hex5-[AEP]HexN-InsP-C23:0/d18:1-Cer	2083.02	EtNP-Hex5-AEP-HexN-Ins-P-23:0/d18:2	109	7
32	[AEP]Hex5-[AEP]HexN-InsP-C40:0(OH)2-Cer	2055.02	AEP-Hex5-AEP-HexN-Ins-P-22:0/d18:0	102	5
33	[AEP]Hex6-HexN-InsP-C23:0/d18:1-Cer	2122.06	Hex6-AEP-HexN-Ins-P-22:4/d18:1	104	6
34	[AEP]Hex6-HexN-InsP-C24:0/d18:2-Cer	2134.06	Hex6-AEP-HexN-Ins-P-23:2/d18:2	105	6
35	[EtNP]Hex5-[AEP]HexN-InsP-C42:2(OH)2-Cer	2096.42	EtNP-Hex5-AEP-HexN-Ins-P-24:2/d18:0	127	3
36	[AEP]Hex5-[AEP]HexN-InsP-C23:0/d18:1-Cer	2067.02	AEP-Hex5-AEP-HexN-Ins-P-22:0/d18:0	102	5
37	[AEP]Hex5-[AEP]HexN-InsP-C24:0/t18:0-Cer	2099.04	AEP-Hex5-AEP-HexN-Ins-P-24:0/t18:1	121	9
38	[EtNP]Hex5-[AEP]HexN-InsP-C24:0/d18:1-Cer	2097.02	EtNP-Hex5-AEP-HexN-Ins-P-24:1/d18:1	127	3
39	[AEP]Hex6-HexN-InsP-C24:0/t18:1-Cer	2154.08	Hex6-AEP-HexN-Ins-P-25:0/t18:2	103	2
40	[AEP]Hex6-HexN-InsP-C23:0/d18:0-Cer	2124.06	Hex6-AEP-HexN-Ins-P-22:2/d18:0	112	11
41	[EtNP]Hex5-[AEP]HexN-InsP-C24:0/t18:0-Cer	2116.04	EtNP-Hex5-AEP-HexN-Ins-P-24:0/t18:0	102	1
42	[AEP]Hex5-[AEP]HexN-InsP-C42:2(OH)2-Cer	2079.02	AEP-Hex5-AEP-HexN-Ins-P-24:0/d18:2	113	3
43	[AEP]Hex5-[AEP]HexN-InsP-C24:0/d18:1-Cer	2081.02	AEP-Hex5-AEP-HexN-Ins-P-24:1/d18:1	113	3
44	[AEP]Hex8-HexN-InsP-C24:0/d18:1-Cer	2460.18	Hex8-AEP-HexN-Ins-P-24:0/t18:0	101	4
45	[EtNP]Hex5-[AEP]HexN-InsP-C24:0/t18:1-Cer	2115.04	EtNP-Hex5-AEP-HexN-Ins-P-24:0/t18:0	102	1
46	Hex6-[AEP]HexN-InsP-C24:0/d18:1-Cer or [AEP]Hex6-HexN-InsP-C24:0/d18:1-Cer	2135.06	Hex6-AEP-HexN-Ins-P-23:2/d18:2	105	6
47	[AEP]Hex5-[AEP]HexN-InsP-C23:0/d18:0-Cer	2069.02	AEP-Hex5-AEP-HexN-Ins-P-22:0/d18:0	102	6
48	[AEP]Hex7-HexN-InsP-C24:0/d18:1-Cer	2298.12	AEP-Hex7-HexN-Ins-P-24:0/d18:1	117	7
49	[AEP]Hex6-[AEP]HexN-InsP-C24:0/d18:1-Cer	2243.1	AEP-Hex6-AEP-HexN-Ins-P-24:0/d18:1	119	8
50	Hex6-[AEP]HexN-InsP-C24:0/d18:0-Cer or [AEP]Hex6-HexN-InsP-C24:0/d18:0-Cer	2137.08	Hex6-AEP-HexN-Ins-P-23:4/d18:0	105	6
51	[AEP]Hex7-HexN-InsP-C24:0/d18:0-Cer	2300.14	Hex7-AEP-HexN-Ins-P-24:0/d18:0	116	8
52	Hex7-HexN-InsP-C24:0/d18:1-Cer	2191.1	Hex7-HexN-Ins-P-18:4/d18:2	17	37
53	[AEP]Hex5-[AEP]HexN-InsPC24:0/d18:0-Cer	2083.06	AEP-Hex5-AEP-HexN-Ins-P-24:0/d18:2	113	3
54	[AEP]Hex8-HexN-InsP-C24:0/d18:0-Cer	2462.18	Hex8-AEP-HexN-Ins-P-24:0/t18:0	101	4
55	[AEP]Hex6-[AEP]HexN-InsP-C24:0/d18:0-Cer	2245.1	AEP-Hex6-AEP-HexN-Ins-P-24:0/d18:0	118	7
56	[EtNP]Hex5-[AEP]HexN-InsPC24:0/d18:0-Cer	2099.04	EtNP-Hex5-AEP-HexN-Ins-P-24:0/d18:1	125	5
57	[AEP]Hex5-HexN-InsP-C24:0/d18:1-Cer	1974.02	Hex5-AEP-HexN-Ins-P-24:0/t18:0	118	6
58	[EtNP]Hex5-[AEP]HexN-InsP-C25:0/d18:1-Cer	2111.04	EtNP-Hex5-AEP-HexN-Ins-P-24:0/t18:0	102	1

59	[AEP]Hex4-HexN-InsP-C42:1(OH)2-Cer	1811.96	AEP-Hex4-AEP-HexN-Ins-P-15:2/d18:0	124	3
60	[AEP]Hex6-HexN-InsP-C25:0/d18:1-Cer	2149.08	Hex6-AEP-HexN-Ins-P-25:3/d18:0	104	7
61	[AEP]Hex4-HexN-InsP-C42:0(OH)2-Cer	1813.98	Hex4-AEP-HexN-Ins-P-22:2/t18:2	107	7
62	[AEP]Hex8-HexN-InsP-C25:0/d18:0-Cer	2498.2	Hex8-AEP-HexN-Ins-P-24:0/t18:0	101	4
63	[AEP]Hex6-HexN-InsP-C25:0/d18:0-Cer	2152.1	Hex6-AEP-HexN-Ins-P-25:3/d18:0	104	7
64	[AEP]Hex7-HexN-InsP-C25:0/d18:0-Cer	2314.14	Hex7-AEP-HexN-Ins-P-25:0/d18:2	114	9
65	[AEP]Hex5-HexN-InsP-C24:0/d18:0-Cer	1976.02	Hex5-AEP-HexN-Ins-P-24:0/t18:0	118	6
66	[AEP]Hex6-[AEP]HexN-InsP-C25:0/d18:1-Cer	2095.06	AEP-Hex6-AEP-HexN-Ins-P-25:0/t18:2	107	27
67	[EtNP]Hex5-[AEP]HexN-InsP-C25:0/d18:0-Cer	2113.06	EtNP-Hex5-AEP-HexN-Ins-P-25:0/t18:2	109	3
68	[AEP]Hex5-[AEP]HexN-InsP-C25:0/d18:0-Cer	2097.06	AEP-Hex5-AEP-HexN-Ins-P-25:0/d18:0	121	9
69	[AEP]Hex6-HexN-InsP-C26:0/d18:1-Cer	2163.1	Hex6-AEP-HexN-Ins-P-26:0/d18:1	103	2
70	[AEP]Hex5-[AEP]HexN-InsP-C26:0/d18:1-Cer	2109.06	AEP-Hex5-AEP-HexN-Ins-P-26:0/t18:1	110	2
71	[EtNP]Hex5-[AEP]HexN-InsP-C26:0/d18:1-Cer	2125.06	EtNP-Hex5-AEP-HexN-Ins-P-26:0/d18:1	109	3
72	[AEP]Hex6-HexN-InsP-C26:0/d18:0-Cer	2165.1	Hex6-AEP-HexN-Ins-P-26:1/d18:0	103	2
73	[EtNP]Hex5-[AEP]HexN-InsP-C26:0/d18:0-Cer	2127.06	EtNP-Hex5-AEP-HexN-Ins-P-26:4/t18:2	103	3
74	[AEP]Hex5-[AEP]HexN-InsP-C26:0/d18:0-Cer	2111.08	AEP-Hex5-AEP-HexN-Ins-P-26:0/t18:1	110	2
75	[AEP]Hex6-HexN-InsP-alkylacyl-C40:0-Gro	2153.1	Hex6-AEP-HexN-Ins-P-AAG-40:0	103	3
76	[EtNP]Hex6-[AEP]HexN-InsP-C24:0/d18:1-Cer	2259.06	EtNP-Hex6-AEP-HexN-Ins-P-25:0/d18:1	107	27
77	[AEP]Hex6-HexN-InsP-C24:0/t18:1-Cer	2153.06	Hex6-AEP-HexN-Ins-P-25:3/d18:0	104	7
78	[EtNP]Hex5-[AEP]HexN-InsP-C24:0/d18:2-Cer	2095.02	EtNP-Hex5-AEP-HexN-Ins-P-24:0/d18:2	127	3

5.2 Discussion

LC-MS/MS is the most efficient tool today for GPI profiling. The amount of data produced in each MS experiment is a major bottleneck in high-throughput GPIomic projects. Efficient computational tools can significantly reduce the amount of time in the analysis of MS data; however, at present the automatic interpretation of these data to annotate GPI structures is absent.

In this work, the identification of GPI structures was achieved by comparing observed fragment masses with theoretical fragment masses. We have used our own theoretical data table that was built starting from a basic set of glycan and lipid components. At present, our method

provides a ranked list of structures for each given parent ion *m/z* value. The overall accuracy of our method was tested with a set of experimentally confirmed structures.

In our method counting the number of matches acts only as a survey to see whether the fragments of a candidate structure are observed within the spectra, thus reducing the set of candidate structures. The current scoring method produces a significant overlap between the scores of correct and incorrect structures identified since a number of isobaric structures are present after counting. The results show that our method has low selectivity, which needs to be improved.

5.3 Future work

Through the work done to date, I have identified several specific issues in GPI structure determination that need to be investigated.

5.3.1 Uncertainty in m/z assignments

To match a spectrum to a structure in the data table, the *m/z* of the parent ion is scanned throughout the data table to find all possible structures with a tolerance of ±2Da of the *m/z* value. Likewise, the match of a predicted fragment value to an observed value in the spectra is assigned within ±1 Da. tolerance.

The choice of these tolerance levels is based on current practice adopted by experimental researchers (Nakayasu et al., 2009) using mass spectrometry. It would be of interest to explore the effect of changing the tolerance levels on the prediction accuracy. Furthermore, incorporating the deviation from the theoretical m/z values in the scoring system (see next section) might provide useful information.

5.3.2 Scoring system

Currently, we use the following simple scoring system:

For each GPI candidate structure that matches the m/z of the parent ion, the algorithm iterates through the fragments to find an occurrence within the DTA with a tolerance of ±1Da of the fragment mass value. For simplicity, we represent a GPI structure as a sequence of blocks as Figure 5.1.

Let m_i = mass of the *i*th block, *i*= 1, 2,…, *l*, where *l* is the number of blocks in the candidate structure. So, we can write

$$M = \sum_{i=1}^{l} m_i$$

where $M =$ parent ion *m/z.* Furthermore, we define the *h*th, *h* = 1,…, *l*, fragment mass as

$$f_h = \sum_{i=1}^{h} m_i$$

Each candidate structure is evaluated by counting the number of fragment matches within the DTA. Suppose there are *k* observed fragment masses, denoted by $s_1, s_2, \ldots, s_k$ listed in the DTA. Let *t* denote the tolerance. For h = 1,…,l, define

$$I_h = \begin{cases} 1 & \text{if } f_h - t \le s_j \le f_h + t \text{ for some } 1 \le j \le k \\ 0 & \text{otherwise} \end{cases}$$

Then $S = \sum_{h=1}^{l} I_h$ represents the score for the candidate structure.

Bayes' theorem can be used for assessing the probability of finding the correct structure assignment within the *n* (we used *n* = 15 in our analysis described before) top-scoring candidates. Given that a structure is among the *n* top-scoring ones, the probability $P(correct|n)$ that it is the correct structure computed as

$$P(correct|n) = \frac{P(n|correct)p(correct)}{P(n|correct)P(correct) + P(n|incorrect)P(incorrect)}$$

where $P(n|correct)$ and $P(n|incorrect)$ are the probabilities of a given correct and respectively incorrect structure being among the top *n* candidates. $P(correct)$ and $P(incorrect)$ are prior probabilities of a correct and incorrect structure, respectively. These probabilities will be calculated from training data of experimentally confirmed structures.

The field of proteomics has produced successful algorithms that calculate a "matching score" and report how closely a given peptide sequence matches the masses identified in the spectra. Examples of these scores were discussed in Chapter 2, but in essence they attempt to solve the Spectrum Matching Problem (Fenyo and Beavis 2003), which can be stated as follows: Given a spectrum *S* and a score threshold *T* for a spectrum-peptide scoring function, find the probability that a random peptide matches the spectrum *S* with score equal to or larger than *T*. Kim et al., (2008) demonstrated that solving the Spectrum Matching Problem is equivalent to computing the False Positive Rates (FPR) of spectral matches, and proposed a method for computing spectral probabilities using generating functions.

In order to improve the specificity of our method for GPIomics experiments we expect that the application of similar techniques used for solving the Spectrum Matching Problem can lead us to a better scoring scheme. Instead of peptides, we need to match monosaccharide fragments to obtain the glycan part of the molecule, followed by a lipid block for the lipid tail (Figure 5).

5.3.3 Probabilistic model for GPI structure

To use a probabilistic scoring system as in Fenyo and Beavis (2003) and Kim et al. (2008), we first need to have an appropriate underlying probabilistic model for GPI structures. Existing scoring schemes for proteomic MS data are all based on a linear sequential model of peptides. GPI structures, on the other hand, contain possible branch points (see Figure 5.1). Also, we must consider the addition of possible lipid structures to the end of the identified glycan. Using

the biosynthetic constraints, one can construct a bivariate Markov model (MM) for GPI structures as follows.

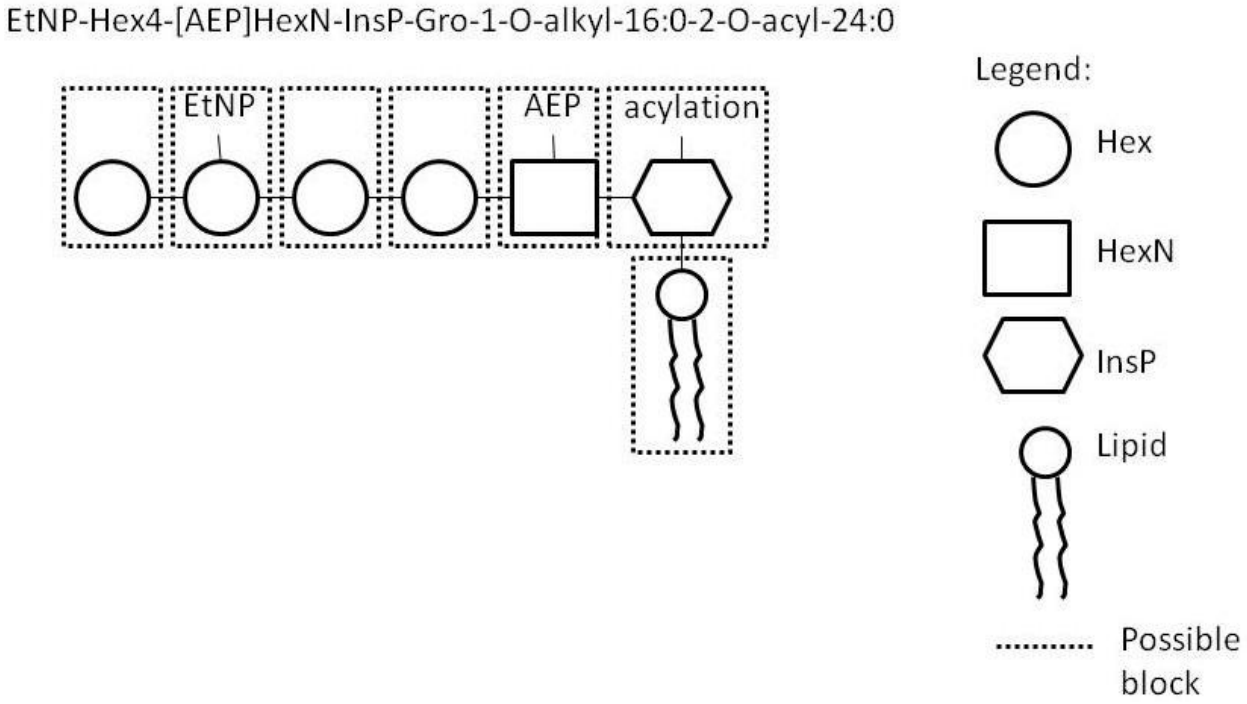


Figure 5.1 Example of states (fragments) for a GPI structure

Consider a set of bivariate states of the form (*m*, *b*) in which *m* represents a monosaccharide, which may be one of Hex, dHex, HexNAc, EtNP, dHex2, NANA, AEP, HexN; *b* represents a possible branching state that may be "null," AEP*, or EtNP*. In addition, there are four possible terminal states represented as (*l*, *t*). The terminal state always consists of the InsP group followed by an acylation, and finally a lipid block that is one of AAG, DAG, lyso and ceramide.

The transition probability matrix and vector of initial probabilities are parameters of the MM that need to be estimated from already known GPI structures. For an *n* state MM, there are $(n+1)(n-1)$ parameters to be estimated. We will try to minimize the number of possible states by applying suitable constraints for glycan branching and lipid composition. Ideally, one would like to construct organism specific models (i.e., we will have one specific model for *T.*

cruzi, one for human, etc.). I will first attempt to construct the model for *T. Cruzi* whose current dataset of GPI structures is most accessible.

References

Chen T, Kao MY, Tepel M, Rush J, Church GM (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **8:** 325-337.

Cooper CA, Joshi HJ, Harrison M J, Wilkins MR, Packer NH (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.* 2003, **31:**511–513.

Dancik V, Addona T, Clauser K, Vath J, Pevzner P (1999) De novo protein sequencing via tandem mass-spectrometry. *Journal of Computational Biology* **6**:327-341.

Domon B, Costello CE (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate* **5:** 397-409.

Eng JK, Fischer B, Grossmann J, MacCoss MJ (2008) A Fast SEQUEST Cross Correlation Algorithm. *Journal of Proteome Research* **7:**4598–4602

Fankhauser N, Mäser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* **21**: 1846–1852.

Ferguson MA (1999) The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contribution of trypanosome research. *Journal of Cell Science* **112**:2799-2809.

Fenyo D, Beavis R (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry* **75**:768–774.

Goldberg D, Sutton-Smith M, Paulson J, Dell A (2005) Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* **5:**865–875.

Ikezawa H (2002) Glycosylphosphatidylinositol (GPI)-anchored proteins. *Biol Pharm Bull* **25:**409-417.

Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, Karlsson NG (2004) Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* **4:**1650–1664.

Kim S, Gupta N, Pevzner PA (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*. **7(8):**3354–3363.

Ma B, Zhang K, Liang C (2005) An effective algorithm for the peptide de novo sequencing from MS/MS spectrum. *Journal of Computer and System Sciences* **70:**418-430.

Maass K, Ranzinger R, Geyer H, Lieth CW, Geyer R (2007) "Glyco-peakfinder" – de novo composition analysis of glycoconjugates. *Proteomics* **7:**4435–4444.

McConville MJ, Ferguson MA (1993) The structure, biosynthesis and function of glycosylated phosphatidylinositols in the parasitic protozoa and higher eukaryotes. *Biochem J* **294**: 305–324

Morelle W, Michalski JC (2005) The mass spectrometric analysis of glycoproteins and their glycan structures. *Current Analytical Chemistry*, **1**:29-57

Nakayasu ES, Yashunsky DV, Nohara LL, Torrecilhas AC, Nikolaev AV, Almeida IC (2009) GPIomics: global analysis of glycosylphosphatidylinositol-anchored molecules of Trypanosoma cruzi. *Mol Syst Biol* **5**:261.

Niemela PS, Castillo S, Sysi-Aho M, Oresic M (2009) Bioinformatics and computational methods for lipidomics. *Journal of Chromatography* **877**:2855‑2862

Nosjean O, Briolay A, Roux B (1997) Mammalian GPI proteins: sorting, membrane residence and functions. *Biochim Biophys Acta*. **1331**:153-86.

Omaetxebarria MJ, Elortza F, Rodriguez-Suarez E, Aloria K, Arizmendi JM, Jensen O N, Matthiesen R (2007) Computational approach for identification and characterization of GPI-anchored peptides in proteomics experiments. *Proteomics* **7**: 1951-1960.

Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching database using mass spectrometry data. *Electrophoresis* **20:**3551-3567.

Vainauskas S, Menon AK (2006) Ethanolamine phosphate linked to the first mannose residue of glycosylphosphatidylinositol (GPI) lipids is a major feature of the GPI structure that is recognized by human GPI transamidase. The journal of biological chemistry **281:**38358–38364.

Yetukuri, L, Katajamaa M, Medina-Gomez G, Seppanen-Laakso T, Vidal- Puig A, Orešič M (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis *BMC Systems Biology* **1:**12.

List of Abbreviations

AAG – alkylacyl glycerol

AEP – aminoethylphosphanate

Cer – ceramide

CID – collision-induced dissociation

Da – dalton

DTA – Peak list

ER – endoplasmic reticulum

EtNP – ethanolamine phosphate

FDR – false discovery rate

GIPL – glycoinositolphospholipid

GLcN – glucosamine

GPI – glycosylphosphatidylinositol

Gro - glycerol

Hex – hexose

HexN – hexosamine

HMM – hidden Markov model

InsP – inositol phosphate

LC-MS – liquid chromatography tandem mass spectrometry

Man – mannose

MM – Markov model

MS – mass spectrometry

MS/MS – tandem mass spectrometry

m/z – mass-to-charge ratio

NANA – N-acetyl neuraminic acid

NN – neural network

PI – phosphatidylinositol

PTM – post-translational modification

Xcorr – cross-correlation score

LC-MS/MS – liquid chromatography tandem mass spectrometry

Vita

Clemente Aguilar received his degree in Veterinary Medicine from the Autonomous University of Ciudad Juarez (UACJ) in Ciudad Juarez, Chihuahua Mexico. While Studying Veterinary Medicine he worked in the area of Technical Support at Delphi Automotive Systems. He has practiced small animal medicine since his degree completion. In 2008 he completed his M. Sc. Degree in Bioinformatics from the University of Texas at El Paso (UTEP), El Paso, Texas. During his Masters, his research focus was on developing a statistical application for finding concentration of inverted repeats in RNA sequences. In January 2009 he joined the Computational Science PhD program at UTEP. He is conducting his thesis under the supervision of Dr. Ming-Ying Leung and Dr. Igor Almeida. The focus on his research is to develop computational methods for mass spectrometry data analysis. After defending his thesis he wishes to continue working in the same program towards his doctoral degree.

Permanent address: 320 Carnival Dr.

El Paso Texas, 79912

This thesis was typed by the author.