A SEMI-AUTOMATIC MODULAR FRAMEWORK FOR QUANTITATIVE GLYCOMICS DATA ANALYSIS

by

CHI XIE

(Under the direction of William S. York)

ABSTRACT

Mass spectrometry (MS) based glycomics techniques are most often used to analyze free oligosaccharides that are chemically or enzymically released from glycoconjugates (e.g., glycoproteins, proteoglycans and glycolipids). However, improved MS methods for analyzing intact glycoconjugates are continuously being developed. Protein-linked *N*-glycans and *O*-glycans are typically released by enzymatic and chemical methods. MS is particularly advantageous for the analysis of complex glycan mixtures containing oligosaccharides that cannot be identified by HPLC alone due to the lack of authentic molecular standards. MS methods that provide accurate molecular mass values can afford direct information regarding the glycosyl residue composition of each glycan in the sample. However, several sets of isomeric glycans, each characterized by a distinct glycosyl composition and molecular mass, are often present in a sample being analyzed. The structure and abundance of each member of such an isomeric set cannot be determined by one-dimensional mass profiling techniques. More sophisticated methods such as multi-stage tandem MS (i.e., MS^n) or combinations of MS with high-resolution separation techniques (e.g., HPLC-MS) are required to resolve the isomeric glycans in such samples.

Although many MS based glycomics tools have been developed from different perspectives, no single tool or platform can fit all the needs for glycomics research. Thus, the integration of glycomics tools to satisfy the needs for a particular glycomics project or research group is a significant challenge. Currently existing software does not include a platform or framework (analogous to the Trans-Proteomic Pipeline (TPP) software for proteomics) that allows users to "personalize" processing of glycomics data according to the input file format and the research aims. Such software, which requires the ability to select and orchestrate the appropriate functional modules designed to accomplish specific aims in the glycomics domain, are required for automatic, high throughput analysis of glycomics data.

We have developed a modular glycomics data processing environment and implemented some of its key components, which support semantically annotated data-exchange formats and workflow engines, enabling the development of interactive modules that perform well-defined data processing tasks. A workflow comprised of modules for data conversion, scaling, extraction, and quantification of rolling-trapping MS data was developed and implemented in this environment. This workflow was applied to experimental data, revealing sources of variation between replicates.

INDEX WORDS: Quantitative Glycomics, Linear Modeling, Simulation Optimization,

Mass Spectrometry, Glycomics Data Analysis

A SEMI-AUTOMATIC MODULAR FRAMEWORK FOR QUANTITATIVE GLYCOMICS DATA ANALYSIS

by

CHI XIE

B.E. Nanjing University of Technology, China, 2000

M.S. Southeast University, China, 2004

M.S. Texas Tech University, Texas, United States 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

A SEMI-AUTOMATIC MODULAR FRAMEWORK FOR QUANTITATIVE GLYCOMICS DATA ANALYSIS

by

CHI XIE

Major Professor: William S. York

Committee: Lance Wells

John A. Miller

Krzysztof J. Kochut

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

December 2012

DEDICATION

To my mother and father, thank you for your unconditional love and support throughout my life. Words cannot express my gratitude for your sacrifice. Without your encouragement, I would not be able to achieve anything. I love you all.

ACKNOWLEDGEMENTS

First and most importantly, I would like to express my gratitude and appreciation to my advisor, Dr. William S. York, for giving me the opportunity to study and work in his group with an extraordinary environment. I am really grateful to his constant guidance, support and encouragement throughout all the years of my graduate school.

I would like to thank my committee, Dr. John A. Miller, Dr. Krzysztof J. Kochut and Dr. Lance Wells, for their invaluable time and expertise. I would also like to thank Dr. Michael Tiemeyer, for his countless help.

Many thanks go to the past and present members of the Complex Carbohydrate Research Center: Jaemin Lim, Rene Ranzinger, Meng Fang and Crissy Dobson for their kindness and assistance. Best wishes in their future endeavors.

Last but not least, I would like to thank my family, my girlfriend and all my friends for their love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 GLYCOMICS

The term of glycomics is defined as the comprehensive study of all glycan structures in a given biological system. The word glycomics is a combination of the prefix "glyco", meaning sugar, and the neologism "-omics", which combines with various prefixes to form nouns with the sense 'all of the specified constituents of a cell, considered collectively or in total[1,2]. The study of glycomics includes a range of scientific disciplines that are applied to study the structure and function of carbohydrates in biological systems[3]. Glycomics is considered to be a subset of the study of glycobiology and describes the scientific investigation and identification of all the glycan molecules which are synthesized within a particular cell, tissue, or organism. Glycomics research mainly aims at creating profiles of currently known glycan structures, as well as finding new glycan structures. Glycomics faces many daunting challenges, including the likelihood that a cell contains a very large number of different glycan structures, which arise by combination of several structural factors, including: the presence of various combinations of different monosaccharide building blocks; different ring forms for the monosaccharides; different sterochemistries (α or β) for the glycosidic bonds connecting the monosaccharides; multiple attachment points that can be used to form glycosidic bonds, leading to the possibility of branched structures with diverse molecular topologies; and different monosaccharide sequences for each possible topology[4,5,6].

1.2 GLYCOSYLATION

In general, the word glycan often refers to the structure of an oligosaccharide, which is a polymer containing a small number (typically less than 10) of component monosaccharides joined by glycosidic bonds. Glycans can also be polysaccharides, which are long chains of carbohydrate monomer units. Glycans can be linear or highly branched. The word glycan is also used to represent the carbohydrate portion of a sugar containing biomolecule, such as a glycoprotein or glycolipid[7]. The most common glycans in eukaryotic organisms are composed of monosaccharide residues that include *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GalNAc), mannose (Man), galactose (Gal), fucose (Fuc) and *N*-acetylneuraminic acid (NeuNAc, sialic acids)[8,9,10,11,12]. Five major classes of protein glycosylation have been defined according to the structure and location of the glycan:

(1) *N*-linked glycosylation, in which glycans are linked to the amide nitrogen atom of an asparagine residue, in the tripeptide consensus sequence – Asn-X-Thr/Ser (where X can be any amino acid except proline). In some rare cases, *N*-glycosylation can occur when the serine or threonine residue is replaced with a cysteine as Asn-X-Cys[13,14,15,16,17].

(2) *O*-linked glycosylation, in which the glycan is linked to the hydroxyl oxygen on the side chain of a serine or threonine. No consensus amino acid sequence exists for *O*-glycosylation sites[18.19,20].

(3) Glycophosphatidylinositol (GPI) anchors, which link lipids to the carboxyl terminus of proteins and serves to attach the modified proteins to cell membranes[21,22,23].

(4) Glycosaminoglycans (GAGs), which are long unbranched polysaccharides containing a

repeating disaccharide unit. GAGs are most often linked to the hydroxy oxygen of serine[24,25]

(5) *C*-mannosylation, in which a mannose is attached to the carbon of tryptophan residues [26,27,28]

In contrast to nucleic acids and proteins, the study of glycans is much more challenging due to the complexity of glycan structures and diversity of the glycosylation process *in vivo.* The biosynthesis of glycans is not template-driven. Rather, the structures of glycans within a cell or organism depend on the activities of glycosyltransferases present in the organelles where glycan synthesis occurs. Glycans play critical roles in various physiological processes such as cell development and differentiation[29,30,31,32,33,34] cell-cell recognition and interaction[35,36], tumor growth and metastasis[37,38,39,40,41,42] and immune recognition and response[43,44,45,46,47.] Recent cancer research shows a correlation between glycosyltransferase expression levels and subsequent changes in glycan structures in several diseases, including inherited diseases[48], the progression of cancer[49] and autoimmune diseases[50,51,52]. Several glycobiologists, including Abbott *et al.*, have pointed out that glycan structures have the potential to play the role of biomarkers to investigate the mechanism of cancer[53].

1.3 MASS SPECTROMETRY (MS)

Since the first studies of mass spectrometry starting in 1912, the principles and progress of mass spectrometry have developed rapidly. The working process of mass spectrometers is summarized as: (1) a sample is prepared and fed into mass spectrometer either manually (e.g., on a MALDI target), by direct infusion or after chromatography (e.g., by LC-MS), (2) ions are generated by ionizing the sample and accelerated to a well-defined kinetic energy, (3) single or

multiple mass analyzers are applied to separate and sort the ions according to their individual *m/z* values (mass to charge ratios), (4) an ion detector is used to count the number of ions as a function of *m/z*, and (5) the processed signal data are output as mass spectra in various formats. In the past several years, mass spectrometry has become more and more important in the study of glycomics to identify and quantify glycans and related biomolecules. Basically, there are two MS strategies: single-MS and MS^n. As implied by its name, single-MS applies only a single round of mass analysis in the instrument in which quasimolecular ions (and sometimes fragment ions) are generated and detected. In contrast, MS^n may include one or more rounds of ion selection, fragmentation and mass analysis, providing more detailed structural information for each selected ion. Both two strategies have been applied broadly to identify glycan structures in biological research[54]:

(1) In general, single-MS usually can offer high throughput, broad mass coverage and direct quantification for the interesting structures. But it often results in weak evidence for structure assignment of biomolecules. Despite that, it is frequently used experimental tool with the advantages of speed, simplicity and coverage.

(2) MS/MS includes an ion selection step, a fragmentation step and an ion analysis step. In the laboratory, these steps are usually accomplished sequentially by a single mass spectrometer instrument. To obtain and maintain vacuum in tandem mass spectrometry, physical connections are established between each individual functional unit, including the ion source, the analyzer and the detector.

(3) MS^n (or more generally tandem MS) refers to MS with multiple rounds of ion selection,

ion fragmentation, and fragment ion analysis. The number of mass-analysis steps is specified by the parameter *n*. MS/MS is a variety of tandem MS in which $n = 2$.

1.3.1 IONIZATION SOURCE

Two widely used ionization techniques are ElectroSpray Ionization (ESI) and Matrix-Assisted Laser Desorption Ionization (MALDI).

In the 1980s, ESI was first introduced by John Bennett Fenn and his collaborators as an ionization method for mass analysis of biological macromolecules in solution[55]. In ESI, a stream of liquid containing the sample is passed through a capillary at high electric potential. The charged solution is sprayed out of the capillary tip at atmospheric pressure and the flow of a nebulizing gas such as dry ditrogen is used to expedite evaporation of the solution and droplet shrinkage. Molecules in the analyte are thus stripped of solvent molecules and gas phase ions are formed[56]. Two major hypotheses have been proposed to explain the final production of gas-phase ions: the ion evaporation model (IEM) and the charged residue model (CRM). IEM suggests that, as the droplet reaches a certain radius, the field strength at the surface of the droplet becomes sufficiently large to trigger the field desorption of solvated ions[57]. In contrast, in CRM, electrospray droplets undergo evaporation and fission cycles, leading to progeny droplets that contain on average one analyte ion or less. The gas-phase ions form after the remaining solvent molecules evaporate, leaving the analyte associated with charged species (such as Na^+) carried by the droplet[58]. Unlike the MALDI technique, ESI usually produces multiply charged ions, which allows the analysis of very large molecules even when the *m/z* detection range of the mass analyzer is relatively small. ESI can be coupled with almost any type of mass analyzer.

Another important advantage of ESI lies in its capability to combine with separation techniques e.g. HPLC to perform mass spectrometry in a real-time mode[59].

The MALDI process involves the ionization of the analyte with the proper combination of laser wavelength and matrix materials[60]. To perform MALDI ionization, the sample is crystallized with a low molecular weight ultraviolet (UV)-absorbing matrix molecule and then irradiated with a brief laser pulse to generate gas-phase ions. The matrix maintains physical separation of analyte molecules to lower desorption energy and absorbs laser light having a wavelength at which the analyte may not have significant absorbance. Absorption of energy from the laser beam causes evaporation and ionization of analytes[61]. Although different lasers can be used in MALDI, UV lasers, which include nitrogen lasers, are most often used. MALDI ionization has an advantage over ESI in the fact that the ionization efficiency in MALDI is less affected by the increase of the mass and size of the molecules[60]. However, its disadvantage lies in the presence of metastable ions, which are formed from ions decomposing during flight. Nevertheless, MALDI primarily produces intact singly charged molecular ions. This method is reasonably tolerant toward the presence of salts, buffers and other additives[62] and has a detection limit in the range from 100 femtomole to 2 picomole. MALDI is considered as one of the most widely used ionization methods for the analysis of biopolymers such as proteins and glycans.

1.3.2 ION ANALYSIS

The goal of ion analysis is to separate the various ions. When charged ions travel through the electronic or magnetic field at the same speed, they will be deflected by the external magnetic force or electric field force. The degree of detection is closely correlated to m/z value and follows

Newton's second law of motion, $F = ma$. The force (F) can be adjusted to allow a specific portion of ions to be selected. Ion analysis is performed by mass analyzer(s) that use different techniques to select or detect ions, e.g., quadrupole ion trap (QIT), Time-Of-Flight (TOF) and Fourier Transform-Ion Cyclotron Resonance (FT-ICR). In FT methods, the analyzer and detector are basically the same unit, and ions are detected by recording an induced AC current rather than DC current produced when the ions strike the detector.

1.3.2.1 QUADRUPOLE TRAP

The quadrupole ion trap (QIT) mass analyzer was developed for the quadrupole mass analyzer by Wolfgang Paul[63, 64] in the early 1950's. This technique led to the development of basic parameters in most current popular instruments in glycomics research labs. Later, Finnigan MAT made breakthroughs in the design of quadrupole mass analyzer in the 1980's[65] that allowed QIT to be used in commercially available instruments. Furthermore, QIT instruments are so flexible that they can couple with ESI and MALDI ionization. In some cases, they can also be coupled with liquid chromatography.

In QIT instruments, the ions produced in the source enter the inlet to reach the trap, where they are trapped by the ring electrode, the entrance and the exit endcap electrodes. Various voltages are applied to these three electrodes in order to form a cavity for the ions to be trapped. The potential of ring electrode radio frequency (RF), which is an AC potential with constant frequency and variable amplitude, forms a 3D quadrupolar potential field inside of the trap. This traps the ions in a stable oscillating trajectory. The exact motion of ions depends on the voltages applied to them as well as their mass-to-charge (m/z) ratios. To detect ions, the potentials are

altered to so that the ion motions are destabilized and this will result in the ions to be ejected through the exit endcap. With gradually changing potentials, the ions are ejected in the order of increasing m/z values. The ejection of ions will form an ion stream, which is can be focused directly onto the detector or processed further (e.g., by ion fragmentation).

1.3.2.2 TOF

In concept, TOF-MS is the simplest method for mass measurement, regardless of hidden complexities in applications within MS instruments[66,67]. TOF is now commonly used for biological analysis applications because of its ability to be coupled with MALDI and ESI and the development of high-resolution and hybrid instruments, e.g. Q-TOF and TOF-TOF. TOF has extreme sensitivity to detect all ions with almost unlimited mass range and high speed of analysis. These characteristics place TOF among most desirable methods for mass analysis.

In TOF, the ions are introduced in two ways: directly from the source of the instrument, or from a previous analyzer as a pulse. The introduction of ions will result in all ions receiving the same initial kinetic energy for them to pass along the field free drift zone. Ions are then separated by their masses, as ions with lower *m/z* travel faster. This process enables the instrument to record all ions at the detector, which accounts for the technique's high sensitivity. In practical application, a kinetic energy distribution for each discrete *m/z* exists, which lowers the resolution by creation of a TOF distribution for each *m/z*[68]. This issue is corrected by the placement of a reflectron at the end of the drift zone[69]. A reflectron consists of a series of electric fields, which repulse the ions back along the flight tube, refocusing the TOF of ions with the same *m/z*.

1.3.2.3 FT-ICR

FT-ICR is perhaps the most complex method in mass analysis, which was first published in the 1950's as a technique to measure small mass differences at very high precision. In the early 1970's, with the application of FT methods[70, 71] by Alan Marshall and Melvin Comisarow[72,73], it became one of the most sensitive and high resolution methods of ion detection.

In the basic FT-MS instrument, ions are generated from the source and then they pass a series of pumping stages at high vacuum, which also keeps increasing. When the ions enter the ion trap, pressures are in the range of [10^{-10} mBar, 10^{-11} mBar] and temperatures are close to absolute zero. The ion trap is located inside a spatially, uniform static superconducting high magnetic field - usually within the range of [4.7 Tesla, 13 Tesla]), which is cooled by liquid helium and liquid nitrogen. When the ions pass into the magnetic field, they are redirected by the Lorentz Force, which is the force on a moving charged particle in a magnetic field, (shown in the following equations) and forced into a circular motion in a plane orthogonal to the field. These ions are therefore prevented from exiting the ion trap cell by the trapping plates at each end.

$$\mathbf{F} = z\mathbf{v} \times \mathbf{B}$$

$$\omega_c = \frac{z\mathbf{B}}{2\pi m}$$

$$\frac{m}{z} = \frac{\mathbf{B}}{2\pi\omega_c}$$

F: the Lorentz Force observed from the ion when it enters the magnetic field

z: the charge on the ion

v: the incident velocity of the ion

B: the magnetic field strength, which is a constant for each instrument

ω_c: the induced cyclotron frequency

m: the mass of the ion

1.3.3 ION DETECTION AND ION TRAP

When positively charged ions enter the electric and magnetic fields, ions will be held in stable orbits for a period of time so that measurements can be performed on them. In FT-ICR, the mass analyzer is combined with the detector and the detected signals generate a complex, convoluted signal of frequency vs. time for all the ions. The resulting data is processed by the instrument, which is usually coupled with a computer. The signal is deconvoluted using Fourier transformation (FT) to generate a deconvoluted frequency vs. intensity spectrum, which is then converted into the mass spectrum. A calibration method is applied for *m/z* error correction. There are two major forms of methods to do the ion detection, omegatron and Fourier transform (FT), both of which apply the cyclotron principle[74], in which the positive ions that flow through a uniform magnetic field follow circular trajectories with a radius that us proportional to momentum and expressed as the function, $r = mv/z\mathbf{B}$. The frequency of rotation can be expressed as a function of mass, $\omega = v/r = z\mathbf{B}/m$.

(1) Omegaton[75]: In this method, the frequency of an oscillator varies in order to bring ions of various masses in tune and increase their momenta to a radius at which a detector is located. Mass can then be directly calculated from frequency since $\omega = v/r = z\mathbf{B}/m$. Resolution of this method can be significantly high when a sufficient magnetic field is used. This method is often used for the analysis of the residual gas within a vacuum and at less than ideal

resolution.

(2) FT[76,77,78,79,80]: In this method, the frequency of the oscillator is swept through the range according to the mass range of interest. Each ion is confined to a specific circular orbit of approximately constant radius with a well-defined frequency. When the oscillator is turned off, an electrode will pick up the radio-frequency radiation of the moving ions. The amplified output can be recorded either directly or after it is mixed with the frequency of a local oscillator. The captured periodic signal is then converted to digital format, which the computer can convert to its frequency spectrum by Fourier transformation. This process is repeated many times in order to enhance accuracy. This method is applied in the design for devices of high resolution. An ion trap is the combination of electric or magnetic fields designed to capture ions in a defined region. Ion traps have a number of scientific applications including mass spectrometry and trapping ions during the manipulation of an ion's quantum state. Ion traps can be divided into two major classes, linear ion trap and orbitrap.

i) In linear ion trap, ions are confined at specific radius by a two-dimensional (2D) RF field and at specific axis by stopping potentials applied to end electrodes. Compared with three-dimensional (3D) Paul traps[81], linear traps have higher injection efficiencies and higher ion storage capacities[82]. In Thermo linear trap quadrupole (LTQ) XL mass spectrometer, the linear ion trap is a square array of hyperbolic rods. In each quadrupole rod section, rods opposite each other in the array are connected electrically, and the four rods of each section can be grouped into two

pairs of two rods each. DC potentials are applied to trap ions in the Z direction while an AC voltage of constant frequency and amplitude is applied to the X and Y rods to trap ions in X and Y directions. Helium is used as dampening gas when trapping to decrease the kinetic energy of the ions to be trapped. The mass spectrum is obtained by scanning the fields at which ions are ejected from the analyzer. LTQ provides higher sensitivity, resolution and mass accuracy at similar mass range than a conventional 3D ion trap and is a robust tool for biochemical analysis[83,84,85].

ii) The Orbitrap mass analyzer is composed of a spindle-shape central electrode surrounded by a pair of bell-shaped outer electrodes[86]. In orbitrap, stable ion trajectories combine rotation around an axial central electrode with harmonic oscillations along it, thus causing ions to cycle around the central electrode in rings[87]. The frequency ω of these harmonic oscillations along the *z*-axis depends only on the ion *m/z* and the field curvature (*k*). The outer electrode of the orbitrap, which has two split halves, detects the induced current from ions of specific *m/z* values moving in rings which oscillate along the central spindle[88]. By Fast Fourier Transformation (FFT) of the image current, the instrument obtains the frequencies of these axial oscillations and therefore the *m/z* of the ions. Nowadays, orbitrap technology is the well accepted standard for accurate mass and is widely used as high-resolution measurement for glycan structure identification. Orbitrap has proven to be a robust mass analyzer for proteomics, glycomics and studies of post

translational modifications.

1.3.4 ELECTRON TRANSFER DISSOCIATION (ETD)

In 1998, McLafferty and his coworkers[89] introduced a unique method for peptide/protein ion fragmentation: electron capture dissociation (ECD), in which low energy electrons are reacted with peptide cations in the magnetic field of a FT-ICR mass spectrometer. With ECD, peptide backbone cleavage is relatively indifferent to either peptide sequence or length. More importantly, ECD does not cleave chemical modifications from peptide and thus labile post-translational modifications remain intact with ECD. However, ECD requires an FT-ICR mass spectrometer, which is the most expensive MS instrument up to now[90].

Also, ECD is a relatively inefficient method since it takes a long time for a scan and therefore it is difficult to couple with a chromatography instrument. As a new method to fragment peptides, ETD was introduced to the proteomics research in recent years[91]. ETD fragments peptide by transferring an electron from a radical anion to a peptide cation and therefore form a radical cation. The ion/ion reaction can be described as follows:

$$(M + nH)^{n+} + A^{-}\bullet \rightarrow (M + nH)^{(n-1)+\bullet} + A \rightarrow \text{ion fragments}$$

Where $A^{-}\bullet$ is the reacting anion.

This reaction results in the fragmentation of peptide backbone. ETD preserves post-translational modifications that are labile during CAD and makes it possible to obtain the sequence information of the peptide including the site of phosphorylation and glycosylation[92]. A large-scale study of ETD performance has been carried out, and it showed that ETD outperforms CAD in the efficiency (completeness) for all charge state greater than two because of its higher

percent fragmentation, which refers to the number of observed fragment ions relative to the number of theoretical fragment ions[93,94,95,96,97,98,99].

1.4 GLYCAN MASS PROFILING

Glycan mass profiling is the quantitative study of the presence of enzymatically or chemically released glycan structures, which are analyzed using various MS methods such as MALDI-TOF-MS. As an example, mass profiling of *N*-linked glycans involves treatment of glycoproteins with endoglycosidases, exoglycosidases or glycoamidases to release *N*-linked glycans from the protein. Permethylation of the glycans followed by mass spectrometry analysis allows identification and quantification for each specific glycan structure to be carried out. Thus, the relative abundance of the various oligosaccharide ions can be delineated from the mass spectral analysis to quantify the glycans present in the sample[100]. When the abundances are normalized such that their sum equals 100%, the results are referred to as "prevalence".

The total ion mapping (TIM) has also been utilized to detect and quantify the prevalence of individual glycans in the total glycan profile[101,102]. In a TIM mode, automated MS and MS/MS spectra are obtained in collection window of a specified width. The *m/z* range from 500 to 2000 is scanned in these successive collection windows, which also have overlap with each other. Due to the immaturity of high-throughput glycan identification software, manual structural assignment of glycans according to individual MS^n spectra is still required. In recent years, more and more automatic or semi-automatic computer programs for glycan structure assignment based on mass spectra such as SimGlycan[103], GlycoWorkbench[104], etc., have been developed to achieve significant saving of time and effort.

1.5 QUANTITATIVE GLYCOMICS METHODS

There are two main types of quantitative methods for mass spectral analysis, absolute quantitative glycomics and relative quantitative glycomics[105,106], which are outlined as follows:

(1) Absolute quantification determines the quantity of each individual glycan in a sample. Within the MS sample, the ion current can represent the abundance of glycan structures, which can be measured by the determining the peak area(s) corresponding to each glycan structure. However, determining the quantity of each individual glycan depends on numerous factors that can change from sample to sample. These include ionization efficiency, sample preparation and matrix effect in the MS instrument.

(2) The techniques of relative quantitative glycomics focus on the determination of how levels of individual glycans change between samples. Strategies for relative quantification using MS analysis address the factors that hinder the absolute quantitative research and generally introduce less error between different experiments by comparing ion intensities of the glycans of interest within the same spectrum.

2. TOOLS FOR GLYCOMICS

With the development of glycomics study and the accumulation of experiment data in the past several years, highly efficient glycomics tools became more and more on the demand of glycobiologists. Glycomics tools are defined as bioinformatics data analysis or processing methods, mostly computer-based, for analyzing glycan structures[107]. With several years' development, many glycomics tools are available, e.g. *GlycoWorkbench*, *CartoonistOne* and *CartoonistTwo*. Many bioinformatics tools focus on identification of glycan structures in

tandem MS, using either database searches or *de novo* methods, which includes *STAT*, *StrOligo*, *GLYCH* and *CartoonistTwo*[108]. Several programs existed for MS^n data file processing, including the well-known *GlycoWorkbench*[109]. In this section, we will introduce different types of glycomics tools and make a brief comparison between them.

2.1 SEMI-AUTOMATIC GLYCOMICS ANALYSIS TOOLS

Automatic glycomics tools are applications or software that can use the data file (e.g. .raw, mzXML, etc.) generated from biological analysis or MS experiment as the input to obtain annotated glycan lists based on data sets and other necessary information associated with the input file. Great progress has been made in the development of automatic annotation platforms or software for mass spectra of carbohydrates and glycoproteins as well as glycolipids. Some semiautomatic glycomics tools exist and are utilized in glycomics research, e.g. *Cartoonist* (for single-MS) and *GlycoWorkbench* (for MS^n), which allows the user to view the spectrum in the data file and search for glycan candidates in the built-in databases based on the *m/z* values and the level of MS. Development of a robust, fully automatic glycomics annotation tool is the goal of several different glycomics groups. Two important semi-automatic glycomics tools will be introduced, *GlycoWorkbench* and *Cartoonist*, and the potential for development of fully automated glycomics tools will be discussed.

2.1.1 *CARTOONIST*

2.1.1.1 *CARTOONIST ONE*

Currently, *CartoonistOne and CartoonistTwo* are among the most well-known glycomics tools to identify glycans by MS[110]. These semiautomatic tool, which provide reasonably accurate

annotation of glycans from the MS data file, have been used by Consortium for Functional Glycomics (CFG) to create glycan profiles for many organisms and tissues.

In general, *Cartoonist* is a useful tool to annotate peaks in MALDI-TOF spectra of permethylated *N*-glycan structures with cartoons which represent the most plausible glycans consistent with the peak masses and the types of glycans being analyzed. The initial step in matching peaks to potential glycans is to construct a table of potential glycans. Single-MS glycan spectra are processed according to the following three steps:

(1) Identify peaks in the spectra file which match the masses and isotope ratios of known glycans compositions. This step uses a built-in table of monosaccharides together with some biosynthetic rules, which include information particular to the species, tissue or other special conditions of each spectrum in the form of scoring demerits.

(2) Generate all biologically plausible N-glycan structure cartoons for each peak. To generate cartoon candidates more correctly, *Cartoonist* also sets up constraints consistent with input MS data files from the experiment. It assumes that the *N*-glycan structures are permethylated, and does this annotation automatically based on peaks within the spectrum itself.

(3) Score the cartoon candidates generated in the last step with the application of expert knowledge and annotate each peak with its most likely cartoon(s). At this step, a table of variables is provides so an expert user can input the values for some critical variables according to the user's knowledge. Then, all plausible annotations will be made and associated with the set of peaks, together with scores indicating the probability that these annotation are correct

As a tool for single-MS, *Cartoonist* is a good choice. It can work as a semiautomatic tool

and give a list of structures that can best explain a spectrum. Since single-MS itself is not a powerful tool to identify glycan without predefined lists of candidate structures, the accuracy of this tool to identify glycans is not that good compared with those semiautomatic glycomics tools based on MS^n [111]. Few semiautomatic tools for single-MS exist and there is still room for progress in the development of fully automatic tools.

2.1.1.2 *CARTOONIST TWO*

CartoonistTwo concentrates on tandem-MS identification of glycan structures. While processing tandem MS data with *CartoonistTwo*, a list of O-linked glycans whose structures best fit the input spectrum is produced. This list includes explanations of the ranking, indicates the most likely structure and provides necessary suggestions regarding which addition experiments are necessary to identify the glycans in the spectrum in a more definitive way. This highly efficient performance comes from its excellent scoring algorithm included within *CartoonistTwo*:

(1) For each spectrum, *CartoonistTwo* processes a sequence of MS spectra with peak significances and recalibration of mass measurements. The union of all the significant peaks observed in all the spectra is taken for further processing. A peak observed more than once is given a greater confidence value. Combining all the peaks from the MS sequence will enable the use of the same identification methods for ETD and ECD spectra, at the loss of some information, which includes the intensities of the peaks in the individual spectra.

(2) In the scoring algorithm, the basic scoring method is first applied, which simply counts the number of spectral peaks above the threshold of intensity explained by fragments of the candidate within the mass tolerance. More advanced scorers are then applied, in which the

missing peaks are often used, which are fragments of a proposed structure that are missing in the input spectrum. The advanced scorers improve the scoring from the base scorer by penalizing for fragments of the candidate not observed, with the penalty set to a small constant times the number of unobserved fragments. The small value of this constant results in unobserved peaks being used only to break ties between topologies that explain equal numbers of peaks.

(3) In the application of scoring algorithm, low-intensity peaks are processed by applying statistical confidence measure based on their intensity and *m/z* values so as to distinguish glycan peaks against noise peaks. To do this, *CartoonistTwo* uses a peak histogram to calculate a p-value to evaluate the chance that a given peak would arise from noise alone. Then a low energy glycan fragmentation model is applied, which is called the shedding model of glycan fragmentation, in which one monosaccharide component is cleaved from the glycan at a time, but the charge will remain in the larger daughter ion. In this way, this tool attempts to predict the peaks that are actually observed in spectra.

Current version of *CartoonistTwo* can determine glycan structures to the level of cartoon. That is, *CartoonistTwo* can only indicate the monosaccharide compositions together with their connection or topology[111]. A limitation of *CartoonistTwo* is that it cannot specify the type of bond. Neither can it distinguish isomeric monosaccharides. Overall, *CartoonistTwo* can generate list of all possible cartoons ranked by score with the most likely cartoon having a unique highest score.

2.1.2 GLYCOWORKBENCH

GlycoWorkbench[109] is a semiautomatic glycomics tool developed by the *EUROCarbDB* for the

MS/MS data processing. Initially the goal of *GlycoWorkbench* is to provide a tool to draw glycan structures rapidly as well as to determine the glycan structures from mass spectrum. Manual annotation of fragment spectra is tedious as it contains a series of many repetitive steps. The automation for these steps is relatively straightforward and this process will reduce the cost of time for sequencing a structure. Like several other semiautomatic glycomics tools, the main task of *GlycoWorkbench* is to evaluate a set of structures proposed by the user. To do this, this tool will compare the theoretical list of fragment masses with the list of peaks derived from the spectrum and then generate a matching list. The following are some beneficial characteristics of *GlycoWorkbench* that decrease the time cost to annotate glycans in the mass spectrum[112]:

(1) This tool has an integrated environment with a user-friendly graphical user interface (GUI). This aspect can provide the user with a nice environment to assemble the glycan structure models. Also in this GUI, the user can calculate the mass of glycans as well as to automatically match their fragments with MS^n data. Furthermore, this tool displays the results, especially for the best candidate after comparison.

(2) *GlycoWorkbench* provides a comprehensive and dynamically increasing set of glycan structural constituents together with an exhaustive collection of fragmentation types as well as a list of annotation options.

The current goal of *GlycoWorkbench* is to provide an automatic tool to support the routine interpretation and assignment of glycan annotation from mass spectrometric data. *GlycoWorkbench* has a built-in visual editor for glycan structures (*GlycanBuilder*) [113]. This graphic editor enables the user to assemble glycan structures with a comprehensive collection of

building blocks. The *GlycoWorkbench*, fragmentation engine is able to calculate a complete list of theoretical fragments due to glycosidic cleavages as well as all the possible ring fragments for every available type of monosaccharide. With the built-in annotation engine, the theoretical list of fragment masses will be compared with the experimental peak-list so as to generate the matching list of fragments and to annotate the glycans structures in the mass spectrum. During this matching process, several types of experimental techniques, various types and quantities of ion adducts, and neutral exchanges are taken into account in order to present a better report with good coverage of glycan structure candidates and better accuracy.

The coordination of *GlycoWorkbench* and *Glyco-Peakfinder* can provide a semiautomatic workflow to annotate glycan structures. This workflow can start from raw mass spectrometric data to generate a report or a full list of assigned glycans. *Glyco-Peakfinder* is used to derive all theoretically possible glycan compositions for a list of *m/z* values[114]. The results can then be applied to search in a structure database or to guide the glycobiologists to identify possible glycan structures. The glycan structure candidates are then transferred into *GlycoWorkbench* where all their possible fragments are generated. Then, in *GlycoWorkbench*, *m/z* values for those fragments are matched against each peak with the desired accuracy. Suitable reports and statistics are provided from *GlycoWorkbench* to compare the quality of annotations for each candidate structure.

Overall, *GlycoWorkbench* is a good glycomics tool, but there are still some aspects that could be improved:

(1) The manual of *GlycoWorkbench* is not well written. This seems to be a trivial problem,

but this will make it difficult for user to understand some functions of this tool. This tool does not provide good technical documentation, which makes it hard for software engineer to understand the code and improve it.

(2) Although *GlycoWorkbench* is only a semiautomatic glycomics tool and its integration with other MS data-processing tools could be improved. With this tool, the user can not open an mzXML or structure file and get the list of glycans or the annotation of the glycans identified in the input file automatically. The user must do some work manually. Users can obtain *m/z* values by looking at the input file or from the glycan structure file built using the platform provided by *GlycanBuilder*. The user can search the possible glycan candidates from three databases (*CFG*, *Carbbank* and *Glycosciences*) for structures consistent with the *m/z* values provided. Users need to decide which *m/z* values to be used for searching databases. This choice is non-trivial for users with limited knowledge of glycobiology.

2.2 LIBRARY-MATCHING-BASED TOOLS

At this time, there are three major databases for complex carbohydrates, *Glycosciences*, *KEGG GLYCAN*, and the database developed by the *CFG*. All three databases are based on the *CarbBank* database developed in the 1990s by the Complex Carbohydrate Research Center (CCRC) at the University of Georgia. Many library-based glycomics tools are designed to leverage information in these databases to search for glycan structure candidates [115].

There are two types of library-based glycomics tools: library-based sequencing tool and library-based matching tool. Library-based sequencing tools usually identify the glycan candidates by matching the mass spectra within MS data file with the theoretical peak list derived

from previously known or identified glycan structures. Similar to the *SEQUEST* method used for protein sequencing, *GlycosidIQ*, a frequently used library-based sequencing tool for glycomics, generates a theoretical peak list for each structure in the database by computing all its theoretical fragments[116]. The best match between the theoretical peak lists and the mass spectra is then derived using a suitable scoring function. Library-based matching tool is a totally different approach from library-based sequencing. Library-based matching compares the unassigned spectra against a library of experimentally determined fragment spectra so as to identify unknown glycans. The use of library-based glycomics tool is currently limited because of the poor availability of reliable standard data. This problem can be reduced by the cooperation of experimental glycobiologists by creating a comprehensive and well-curated public collection of assigned MS^n spectra from pure glycans.

2.3 ONLINE TOOLS ASSISTING MS SPECTRA INTERPRETATION

To help glycomics research in an immediate way, many online tools offer some aid to glycobiologists in finding to find the glycan compositions from MS data file. Of course, they are just tools that require judgments by the user.

GlycanMass[117] is a simple web-based tool within *ExPASy*, which can help to calculate the monoisotopic or average mass for a given oligosaccharide composition based on the glycan structure and other information, e.g. whether the oligosaccharide is underivatised, permethylated or peracetylated.

GlycoMod, which is considered as an up-graded version of *GlycanMass*, is a web-based tool designed to search all possible compositions of a glycan structure from its experimentally

determined masses. With the input values of variables from the user, this tool will search for all combinations of composition matching the input mass. This is a nice tool to predict the composition of any glycoprotein-derived oligosaccharide with the consideration of whether this glycan structure is underivatized, methylated or acetylated monosaccharides, and which derivatized reducing terminus this structure is associated with[118].

The *GlycoFragment* allows the generation of all theoretically possible fragments of oligosaccharides. The extended IUPAC nomenclature is used to input structures and several forms of derivatization and substitution of the reducing end have been implemented[119].

2.4 *de novo* APPROACHES

Including *GlycoWorkbench*, several glycomics tools based on *de novo* approaches have been developed to process data from MS^n fragmentation experiments so as to derive the complete structure[104].

2.4.1 *STAT*

STAT, which is a saccharide topology analysis tool, generates all the possible structural topologies from a composition selected by the user among those compatible with the precursor mass. The candidate structures are then compared with the given peak list after the matching list of structures are ranked according to their score[120].

In *STAT*, all possible compositions of the input glycans are first determined by the input information, which include the component monomers, total mass of the glycan structure, mass of the charge carrier, bound threshold of error allowed for the mass, and whether the oligosaccharide contains a reducing terminus. Then a version of "knapsack algorithm"[121] is

applied to calculate all possible combinations of selected monomers, which sum to the adjusted total mass with the application of the specified error bound threshold, and the calculations are based on exactly the same algorithm but twice for each possible composition, once corresponding to the oligomers with reducing termini and once with so called “anhydro” structures.

To calculate all possible topologies containing all substructures implied by the product ions, each possible topology is calculated based on the given composition. This step is then followed by a computation for the mathematical restrictions corresponding to each substructure in order to eliminate those topologies which cannot satisfy all the requirements, in which the generation of free trees, multiset permutations and removal duplicates are carried out by using a tree isomorphism algorithm. Then each candidate in the list will be evaluated by a numerical rating score based on the likelihood to the correct topology, higher scores corresponding to structure with better likelihood. This tool is computationally efficient and for an oligosaccharide having seven or eight monosaccharide units, a list of all the possible saccharide topologies can be generated within several seconds.

At its current stage, the *STAT* software is a semi-automated approach for glycomics data analysis based on MS^n data and, although the algorithm efficiency is still under improvement, it can provided a good starting point for the development of completely automated tool for the structure analysis of oligo- saccharides as well as polysaccharides. However, this software does not yet take the permethylated oligosaccharides into consideration. It would be better if it allowed the user to personalize the analysis by entering structure information, e.g. number and type of core structures.

2.4.2 *OSCAR*

Oligosaccharide Subtree Constraint Algorithm (*Oscar*) was developed as an algorithm for assigning oligosaccharide topology from MS^n Data[122]. *Oscar* translates the selection of MS^n ion fragmentation pathways into possible oligosaccharide topologies. Without presumed biosynthetic constraints or any external comparisons and guided by a series of logical constraints, this approach is applied to provide molecular topology. For example, *Oscar* can be applied to a series of permethylated oligomers so that topologies will be assigned in several seconds. Like *STAT*, *Oscar* generates candidate structures from an estimated composition. But it uses the information contained in fragmentation pathways of permethylated oligosaccharides so as to restrict the set of possible results, which is quite different from *STAT*.

As for the algorithm, the input is a series of fragmentation pathways within in the MS^n data sets obtained by analysis of permethylated oligosaccharides. With the aid of a composition database, *Oscar* tries to identify compositions consistent with fragmentation pathways and that match logical structural constraints. A set of structures that satisfies all the selected pathways will be provided as output. Based on the structural information as well as pathways, the output can converge to a single topology. This algorithm is considered to be a *de novo* approach as it does not search through a database containing the known structures.

Although *Oscar* has great potential to facilitate glycan topology analysis, there are some aspects that need to be improved. These include better support for resolving structural isomers in order to facilitate automated identification of ions that are inconsistent with the candidate structure. Another aspect that needs improvement is its standard library, which contains known

glycan substructures with experimentally observed spectra. The library should be sufficiently populated such that data set comparison provides a way to assign monosaccharides to substructural motifs in unknown glycans.

2.4.3 *STROLIGO*

StrOligo is a tool for the structure assignment of complex glycan structures from glycoproteins using tandem mass spectrometry. In *StrOligo*, the differences between fragment masses are used to estimate the loss of known moieties and to produce a candidate composition for the precursor ion. Given the estimated composition, a set of structures is generated by applying biosynthetic rules specific to mammalian *N*-glycans[123].

In *StrOligo*, a relationship tree is built to account for each observed loss of a monosaccharide moiety. The relationship tree is then analyzed and possible glycan structures, associated with adducts and fragment ion types, are proposed together with a score assigned to each of the proposed glycan structures to show the confidence of the annotation. Then this software attempts to decide which structure candidate is the best fit for the corresponding peaks with the experimental data set. The usefulness of the algorithm has been demonstrated by assigning structures to several *N*-glycans from MS/MS data.

StrOligo is able to assign most probable structures to several oligosaccharides directly from their tandem mass spectra, which represents a considerably more efficient approach than manual interpretation. *StrOligo* can also be used to identify structures in mass profiling studies by providing the goodness of fit with experimental tandem mass spectra. However, in some cases, only a portion of glycan structures can be analyzed due to limitations of the MS instrument,

especially in the negative mode[123].

2.4.4 *GLYCH*

GLYCH, a novel algorithm for glycan characterization using MS/MS data file, is derived from *de novo* peptide sequencing tools by allowing branches in the polymer structure. This algorithm first applies a scoring scheme to identify potential bond linkages between monosaccharides, based on the appearance pattern of cross-ring ions to generate a series of ions from the leaves of the glycan tree structure. A dynamic programming algorithm is used to determine the most probable sets of oligosaccharide structures from the tandem mass spectrum data. The complete list of candidate structures is then generated using the ions which rank highly according to their scores in double cleavages[124].

GLYCH is a useful tool for the glycan assignment to MS/MS data set, however, the current version of *GLYCH* prefers linear structures to branching structures and this issue has to be addressed by the future improvement of the *GLYCH* scoring algorithm. Moreover, *GLYCH* currently can only be applied to the characterization of the oligosaccharides. Identification of glycoproteins and determination of the modification site is taken into consideration but still under development. This difficult problem is a challenge for the development of more advanced methods.

2.5 HORN-TRANSFORM

The Horn-Transform was originally developed as an automated algorithm for the proteomics research, in which a peak-picking method is applied for isotopic clustering in the mass spectrum and then the resulting clusters are subjected to primary charge determination, calculation of the

abundance of the monoisotopic peak and least-square fitting. With the availability of a predicted protein sequence, the resulting set of *m/z* values can be assigned as a data fragment. This approach also proposed a useful procedure for the calculation of signal-to-noise to determine baseline and noise width[125].

At the implementation level, the Horn-Transform algorithm requires three parameters as the input, including the expected range of *m/z* in the spectral segment, the initial value of single-to-noise threshold for the fitting of isotopic peak set as well as the expected charge range within the spectral fragments[125]. Also the user can define the minimal reliability value needed to do the least-square fit. Then this algorithm automatically performs the following procedure:

(1) identification of the isotopic peak cluster from the ion fragment

(2) determination of the charge by measuring the spacings of peaks

(3) calculation of the expected set of relative abundances for the above peaks for a specific elemental formula

(4) fit isotopologue abundances so that observed pattern best matches a given elemental formula in order to identify monoisotopic mass

(5) application of the expected mass together with other structural information, e.g. methylation, end structure, adduct, etc., for the characterization of the isotopic peak cluster

The above procedure will repeat until the characterizations of all possible isotopic clusters in the *m/z* range have been finished.

3. CURRENT CHALLENGES FOR FUTURE DEVELOPMENTS OF GLYCOMICS TOOLS

3.1 CHALLENGES IN GLYCOMICS DATABASES

Currently, there are several well-established glycomics databases available and many library-based glycomics tools are built based on them, but the establishment of comprehensive and well-curated databases still has a long way to go. Furthermore, data organization methods should be re-evaluated and probably re-designed. The procedures by which glycobiologists send their experimental data for a certain glycan from a specific environment to the generally accepted databases has to be improve to make it work more efficiently. Currently existing ways to maintain the glycomics databases and the interfaces that allow the saved data to be extracted from database are not adequate. Procedures that are as robust as those routinely used in genomics and proteomics should be established. However, this situation has not be solved for the following reasons[126]:

(1) The input of glycan structures is much more complicated than those in genomics and proteomics. No well-established GUI or gateway is available for untrained glycobiological researchers to use intuitively.

(2) Currently, several international institutions like NCBI or EMBL-EBI support structural databases. However, existence of these resources does not guarantee the sustainability of deposited glycomics data since those institutes do not support glycan structure databases. Furthermore, no general agreement about the data type has been set for glycomics database.

(3) Publishers in glycomics-related journals do not require the glycomics researchers to

deposit their primary data into the database prior to publication, which is quite different from current the situation in genomics and proteomics. This situation needs to be changed by requiring the publication authors to deposit their data together with the quantitative conclusions and other related information into a well-designed database. This will facilitate the review process. Of course, these data should be universally formatted in an accepted format, e.g. mzXML for raw file and xml for resulting file.

(4) There is no high quality measurement for the annotation of glycomics data generated in the lab. The newly generated data must be evaluated to ensure the quality of the data as well as the quality of glycan structure annotation before deposit into database. Measures of data quality need to be established for glycomics as previously defined for genomics data quality control.

3.2 CHALLENGES IN AUTOMATIC TOOL FOR GLYCOMICS

Development and creation of automatic glycomics tools to interpret mass spectrum for glycans is a very active research area in glycomics and several new tools have been developed in recent years. However, currently there is no general solution for the automated identification and quantification of glycans. The ability to identify proteins rapidly and reliably from mass spectra is one of the driving forces for the development of proteomics. However, in glycomics, the unfortunate situation is that many algorithms and services are either not publicly available and/or run only on a special hardware platform.

Rapid and reliable identification of glycan components from mass spectra is currently a bottleneck for many glycomics projects. Many groups are now trying to develop algorithms or tools to solve this problem. The current availability of glycomics or more specifically, glycan

analysis tools, suggests that the automation of high-quality glycan identification and quantification from MS data is still in its infancy with only a few methods being available. The sensitivity and reliability of tools currently under development are still under evaluation.

Currently, there is no completely automated tool that can automatically use data files in standard formats (such as mzXML) as input and give the list of annotated glycans as well as quantitative analysis result as output. All available tools are now semi-automatic, e.g. *GlycoWorkbench*. This requires the user to have at least some basic knowledge in chemistry and biochemistry to identify peaks that are candidates for annotation as a specific glycan. To solve this problem, a convenient way to make lists of the most likely *m/z* values for glycans in each specific cell type would be very useful. Glycomics tools that provide an intuitive interface for the user to input the values of some experimental variables or parameters will reduce effort and improve the quality of glycan identification results.

3.3 CHALLENGES IN INTEGRATION OF GLYCOMICS TOOLS

Although there are many available glycomics tools, no single tool can fit all the needs for glycomics research. Integration of existing glycomics tools to satisfy the needs for a certain project or a certain group is not trivial. Current glycomics tools can accomplish tasks in the following four areas:

(1) Convert raw file format to mzXML format, which is an open data format for storage and exchange of mass spectroscopy data, developed at the SPC/Institute for Systems Biology [127].

(2) View mzXML files and convert it into other file format or list of peaks.

(3) Detect peaks from profile data file.

(4) Deisotope peaks to transform isotopic ion clusters into single entities, each represented by a single *m/z* value and a single abundance value.

However no tool offers a platform that allows users to design the glycomics workflow they need according to the input file format and their needs for glycomics data processing. This would require a set of interactive functional modules that can be combined according to the aims of a specific experimental setup. Tools should be developed to facilitate the selection and orchestration of such modules to implement workflows or pipelines that automatically process data for each type of experimental setup.

Such a system requires a standardized file formats that permit reading, writing and viewing of the data. It seems that mzXML should be considered as a candidate since it is a good extension of XML for mass spectra data with many translators to convert files from other data formats, and many open-source parsers to read it. Secondly, each module should be able to accomplish a single well-defined task rather than an arbitrary combination of many tasks designed to accommodate a specific experimental protocol. Third, a good platform or GUI should facilitate use by biologists rather than computer scientists. A GUI should integrate the data processing enabling the user to choose functional components so as to form a "personalized" workflow to fit their glycomics analysis. The resulting environment should have good documentation and user guide that is accessible to scientists without an extensive computer science background.

3.4 DEVELOPMENT IN THE FUTURE

The development of bioinformatics tools and databases has been very rapid in recent years. But

this field is still in its infancy compared to genomics and proteomics. In the area of glycomics, many high-throughput research projects require the development of efficient automatic methods for the identification and characterization of glycan structures. Moreover, several major glycomics projects are being carried out and many glycomics databases (CFG, KEGG Glycan, Glycosciences, etc.) are developing. Hopefully these initiatives will provide well-organized trustworthy glycomics data in the near future. Furthermore, these databases will provide a good source for glycomics data mining and analysis [128]. Progress in the development of automated MS data analysis tools as well as glycomics databases, will increase the pace at which glycomics will reaches a higher level sophistication and reaches its potential to solve fundamental problems of biology.

4. SIMULATION

Investigations aimed at gaining an in-depth understanding of how biological and chemical systems function in real world and are often based on simulation with/without explicit models. System models for such cases can be classified into physical and mathematical models as following:

(1) Physical model: often uses to build a representation of the targeted system in real life.

(2) Mathematical model: usually applies symbolic notations and mathematical equations.

Simulation is usually considered as a particular type of mathematical model it can be further divided into different classes according to the following attributes:

(i) deterministic or stochastic

(ii) discrete or continuous

(iii) static or dynamic.

In simulation research, the difficulty often lies in the transformation or implementation of a conceptual model to generate an executable simulation program that bridges the huge gap between them. A model is often expressed in natural language and/or mathematics while an executable simulation program is often written in defined programming languages[129].

5. OPTIMIZATION METHOD

Simulation, which is an approximation to the real world, is used when the complexity of the problem makes it impossible to enumerate all possible scenarios by brute force to identify the best solution. Therefore, optimization techniques are necessary to guide the simulation toward an acceptable solution to the real life problem. On the other hand, without the aid of simulation, some real-world problems are too complex to model by explicit mathematical functions. This makes some traditional optimization techniques (gradient-based approaches and random walk methods) impractical[130]. This major dilemma is a barrier to making simulation that approximates real world scenarios that provide useable information. Optimization of simulation methods addresses this problem with the combination of those two approaches.

In general, continuous parameter estimation can be classified into gradient based[131,132] and non-gradient based[133,134]. Gradient-based techniques and stochastic approximation are the two of the most often used optimization algorithms for the problems of continuous parameter estimation. Non-gradient approaches include the Nelder-Mead (simplex) method[135], the Rosenbrock methods[136], and the Hooke and Jeeves method[137], which are gradient free and can efficiently solve target problems for which calculating the explicit derivatives is costly; but

calculating the fitness function itself is efficient.

For the scenario when it is possible to enumerate the possible combinations of the parameters, statistical selection methods, which include subset selection, indifference-zone ranking and selection, multiple comparison procedures, can be utilized to get a small feasible region via subset selection or the optimal solution itself. However, in most cases, the search space is so huge that the enumeration and evaluation of each candidate solution is impossible. In such cases, some kind of random walk methods can be used to overcome the combinatorial explosion and the limitations of computer power. Those methods include simulated annealing and Tabu search.

The following is a description of two optimization approaches that illustrate the classification of gradient and non-gradient methods.

5.1 CONJUGATE GRADIENT METHOD

In mathematics, the conjugate gradient method[137] is an algorithm for the numerical solution of linear equations, whose matrix is symmetric and positive-definite, according to its name. The conjugate gradient method is an iterative method, so it can be applied to sparse systems that are too large to be handled by direct methods. Such systems often arise when numerically solving partial differential equations.

The conjugate gradient method can also be used to solve unconstrained optimization problems such as energy minimization[138]. The biconjugate gradient method is an approach to provide a generalization to non-symmetric matrices. Various nonlinear conjugate gradient methods seek minima of nonlinear equations.

The implementation of this method is based on the system of linear equations, denoted as **Ax**=**b**, where **A** is an *n*-by-*n* symmetric, positive definite matrix, **x** is the solution vector and **b** is the known vector. To calculate the solution **x**, starting with the initial guess of vector $\mathbf{x_0}$, the first step for this approach is the calculation of the residual vector $\mathbf{r_0}$ associated with $\mathbf{x_0}$ from the formula $\mathbf{r_0} = \mathbf{b} - \mathbf{Ax_0}$, where $\mathbf{r_0}$ serves as the initial search direction $\mathbf{p_0}$ as well as the scalar α_0:

$$\alpha_0 = \frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}{\mathbf{p}_0^{\mathrm{T}}\mathbf{A}\mathbf{p}_0}$$

The $\mathbf{x_1}$, the scalar $\boldsymbol{\beta_0}$, search direction $\mathbf{p_1}$ as well as $\mathbf{x_2}$, which is a partially optimized approximation of the solution to the system, are computed using the following formulas:

$$\mathbf{x}_1 = \mathbf{x}_0 + \alpha_0 \mathbf{p}_0$$

$$\beta_0 = \frac{\mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1}{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}$$

$$\mathbf{p}_1 = \mathbf{r}_1 + \beta_0 \mathbf{p}_0$$

$$\alpha_1 = \frac{\mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1}{\mathbf{p}_1^{\mathrm{T}}\mathbf{A}\mathbf{p}_1}$$

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1 \mathbf{p}_1$$

5.2 NELDER-MEAD ALGORITHM

The Nelder–Mead method[139], also called simplex method, is a commonly used linear optimization technique proposed by John Nelder & Roger Mead[135] in 1965. It is a well-defined numerical method for twice differentiable and unimodal problems by minimization of a set of objective functions in a many-dimensional space. The Nelder–Mead method applies the

concept of a simplex, a special polytope of $N + 1$ vertices in N dimensions, and approximates a local optimum of a problem with N variables when the objective functions vary smoothly from one set of variables to another set.

For example, in the simulation of monoisotopic peak in MS data, the simulator has to choose the peak-width and delta that are consistent with the experimental data set, as described in next chapter, which are interdependent and it is difficult to investigate the impact of changes from any specific variable. The Nelder–Mead method can be applied to generate trial calculation results and the fitness can be tested after each round of simulation trials to determine whether an optimized set of parameters has been achieved.

In our application, Nelder–Mead generates a new test simulation as well as the fitness measure by extrapolating the behavior of the objective function. Here we use the Pearson Correlation function of linear regression as the objective function, and trial values of variables arranged as a simplex. In this two-dimensional problem, the simplex is a triangle whose vertices are the coordinate of three trial values for σ and peak width (pw), which will be described in details in next chapter. The algorithm then chooses to replace one of these test sets of variables with the new sets and so the optimization of fitness can progress. The simplest way is to remove the worst set of variable and use a new variable set which comes from inversion through the centroid of the remaining N sets. This loop will continue following the line until this set of variable is not better than the best current set. Then we are able to enter a local valley to limit the simplex towards a better target set of values for variables.

5.3 NEWTON-RAPHSON METHOD

In numerical analysis, the Newton–Raphson method[140], also known as Newton's method, is a mathematical optimization method to find successively better approximations to the roots of a real-world objective function. The general idea of the Newton- Raphson method[141] is, starting with an initial guess of variable reasonably close to the true root, the first derivative of the given objective function is approximated by its tangent line, to calculate the *x*-intercept of this tangent line, which will be considered as a better approximation to the function's root than the originally guessed value of *x*, and the method can be iterated until *x* reaches its optimized value, where the function reaches zero. When the Newton-Raphson method is applied to the first derivative of a function, optimization corresponds to a minmum or maximum of the function.

The Newton-Raphson method can be implemented as following for the case of one variable: Given an objective function *f(x)* defined over the real variable *x*, and its derivative *f '(x)*, a first guess x_0 for a root of the function *f* will be made. Based on the provided function, a better approximation x_1 is generate by the function

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

This calculation is looped until a sufficiently accurate value of *x* is generated, which is the optimized value for *x*. (i.e., *f*(*x*) = 0).

REFERENCE

1. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; and Haslam, S. Journal of Proteome Research 7:1650–1659, (2008)
2. Ceroni, A., Joshi, H; Maa, K.; Ranzinger, R.; and von der Lieth, C-W. Glycoscience, 10.5: 2219-2240, (2008)
3. Ceroni, A.; Dell, A; and Haslam, S. Source Code for Biology and Medicine, 2:3-15, (2007)
4. Collins, B. E. & Paulson, J. C. Cell surface biology mediated by low affinity multivalent protein-glycan interactions. Curr Opin Chem Biol 8, 617-625, doi:10.1016/j.cbpa.2004.10.004 (2004).
5. Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C. & Sasisekharan, R. Glycomics: an integrated systems approach to structure-function relationships of glycans. Nat Methods 2, 817-824, doi:10.1038/nmeth807 (2005).
6. Raman, R., Sasisekharan, V. & Sasisekharan, R. Structural insights into biological roles of protein-glycosaminoglycan interactions. Chemistry & biology 12, 267-277, doi:10.1016/j.chembiol.2004.11.020 (2005).
7. Aoki-Kinoshita K.F. PLoS Computational Biology, 4(5): 99-112, (2008)
8. Butor, C., Diaz, S. & Varki, A. High level O-acetylation of sialic acids on Nlinked oligosaccharides of rat liver membranes. Differential subcellular distribution of 7- and 9-O-acetyl groups and of enzymes involved in their regulation. J Biol Chem 268, 10197-10206 (1993).
9. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3,

97-130 (1993).

10. Morelle, W., Faid, V., Chirat, F. & Michalski, J. C. Analysis of N- and O-linked glycans from glycoproteins using MALDI-TOF mass spectrometry. Methods Mol Biol 534, 5-21, doi:10.1007/978-1-59745-022-5_1 (2009).
11. Pabst, M. & Altmann, F. Glycan analysis by modern instrumental methods. Proteomics 11, 631-643, doi:10.1002/pmic.201000517 (2011).
12. Aoki-Kinoshita, K. F. An introduction to bioinformatics for glycomics research. PLoS Comput Biol 4, e1000075, doi:10.1371/journal.pcbi.1000075 (2008).
13. Sato, C. et al. Characterization of the N-oligosaccharides attached to the atypical Asn-X-Cys sequence of recombinant human epidermal growth factor receptor. Journal of biochemistry 127, 65-72 (2000).
14. Vance, B. A., Wu, W., Ribaudo, R. K., Segal, D. M. & Kearse, K. P. Multiple dimeric forms of human CD69 result from differential addition of N-glycans to typical (Asn-X-Ser/Thr) and atypical (Asn-X-cys) glycosylation motifs. J Biol Chem 272, 23117-23122 (1997).
15. Gil, G. C., Velander, W. H. & Van Cott, K. E. N-glycosylation microheterogeneity and site occupancy of an Asn-X-Cys sequon in plasmaderived and recombinant protein C. Proteomics 9, 2555-2567, doi:10.1002/pmic.200800775 (2009).
16. Chi, Y. H. et al. N-glycosylation at non-canonical Asn-X-Cys sequence of an insect recombinant cathepsin B-like counter-defense protein. Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology 156, 40-47, doi:10.1016/j.cbpb.2010.01.017 (2010).

17. Matsui, T. et al. N-glycosylation at noncanonical Asn-X-Cys sequences in plant cells. Glycobiology 21, 994-999, doi:10.1093/glycob/cwq198 (2011).

18. Gerken, T. A., Owens, C. L. & Pasumarthy, M. Site-specific core 1 Oglycosylation pattern of the porcine submaxillary gland mucin tandem repeat.Evidence for the modulation of glycan length by peptide sequence. J Biol Chem 273, 26580-26588 (1998).

19. Asada, M. et al. The AATPAP sequence is a very efficient signal for Oglycosylation in CHO cells. Glycoconj J 16, 321-326 (1999).

20. Schlummer, S. et al. Influence of serine O-glycosylation or O-phosphorylation close to the vJun nuclear localisation sequence on nuclear import. Chembiochem :a European journal of chemical biology 7, 88-97, doi:10.1002/cbic.200500212 (2006).

21. de Lederkremer, R. M., Lima, C., Ramirez, M. I. & Casal, O. L. Structural features of the lipopeptidophosphoglycan from Trypanosoma cruzi common with the glycophosphatidylinositol anchors. European journal of biochemistry / FEBS 192, 337-345 (1990).

22. Puoti, A., Desponds, C., Fankhauser, C. & Conzelmann, A. Characterization of glycophospholipid intermediate in the biosynthesis of glycophosphatidylinositol anchors accumulating in the Thy-1-negative lymphoma line SIA-b. J Biol Chem 266, 21051-21059 (1991).

23. Puoti, A. & Conzelmann, A. Structural characterization of free glycolipids which are potential precursors for glycophosphatidylinositol anchors in mouse thymoma cell lines. J Biol Chem 267, 22673-22680 (1992).

24. Glowacki, A., Kozma, E. M., Olczyk, K. & Kucharz, E. J. [Glycosaminoglycans--structure and function]. Postepy biochemii 41, 139-148 (1995).

25. Sasisekharan, R., Raman, R. & Prabhakar, V. Glycomics approach to structurefunction relationships of glycosaminoglycans. Annual review of biomedical engineering 8, 181-231, doi:10.1146/annurev.bioeng.8.061505.095745 (2006).

26. Krieg, J. et al. C-Mannosylation of human RNase 2 is an intracellular process performed by a variety of cultured cells. J Biol Chem 272, 26687-26692 (1997).

27. Gonzalez de Peredo, A. et al. C-mannosylation and o-fucosylation of thrombospondin type 1 repeats. Mol Cell Proteomics 1, 11-18 (2002).

28. Perez-Vilar, J., Randell, S. H. & Boucher, R. C. C-Mannosylation of MUC5AC and MUC5B Cys subdomains. Glycobiology 14, 325-337, doi:10.1093/glycob/cwh041 (2004)

29. Lin, X. Functions of heparan sulfate proteoglycans in cell signaling during development. Development 131, 6009-6021, doi:131/24/6009 [pii] 10.1242/dev.01522 (2004).

30. Inatani, M., Irie, F., Plump, A. S., Tessier-Lavigne, M. & Yamaguchi, Y. Mammalian brain morphogenesis and midline axon guidance require heparin sulfate. Science 302, 1044-1046, doi:10.1126/science.1090497 302/5647/1044 [pii] (2003).

31. Hwang, H. Y., Olson, S. K., Esko, J. D. & Horvitz, H. R. Caenorhabditis elegans early embryogenesis and vulval morphogenesis require chondroitin biosynthesis. Nature 423, 439-443, doi:10.1038/nature01634 nature01634 [pii] (2003).

32. Lowe, J. B. & Marth, J. D. A genetic approach to Mammalian glycan function. Annu Rev Biochem 72, 643-691, doi:10.1146/annurev.biochem. 72.121801.161809 121801.161809 [pii]

(2003).

33. Haltiwanger, R. S. & Lowe, J. B. Role of glycosylation in development. Annu Rev Biochem 73, 491-537, doi:10.1146/annurev.biochem.73.011303.074043 (2004).
34. Cipollo, J. F., Awad, A. M., Costello, C. E. & Hirschberg, C. B. N-Glycans of Caenorhabditis elegans are specific to developmental stages. J Biol Chem 280, 26063-26072, doi:M503828200 [pii] 10.1074/jbc.M503828200 (2005).
35. Collins, B. E. & Paulson, J. C. Cell surface biology mediated by low affinity multivalent protein-glycan interactions. Curr Opin Chem Biol 8, 617-625,doi:10.1016/j.cbpa.2004.10.004 (2004).
36. Crocker, P. R. Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling. Curr Opin Struct Biol 12, 609-615, doi:S0959440X02003755 [pii] (2002).
37. Ishida, H. et al. A novel beta1,3-N-acetylglucosaminyltransferase (beta3Gn-T8), which synthesizes poly-N-acetyllactosamine, is dramatically upregulated in colon cancer. FEBS Lett 579, 71-78, doi:S0014-5793(04)01427-9 [pii] 10.1016/j.febslet.2004.11.037 (2005).
38. Fuster, M. M., Brown, J. R., Wang, L. & Esko, J. D. A disaccharide precursor of sialyl Lewis X inhibits metastatic potential of tumor cells. Cancer Res 63, 2775-2781 (2003).
39. Liu, D., Shriver, Z., Venkataraman, G., El Shabrawi, Y. & Sasisekharan, R. Tumor cell surface heparan sulfate as cryptic promoters or inhibitors of tumor growth and metastasis. Proc Natl Acad Sci U S A 99, 568-573, doi:10.1073/pnas.012578299 99/2/568 [pii] (2002).
40. Dube, D. H. & Bertozzi, C. R. Glycans in cancer and inflammation--potential for

therapeutics and diagnostics. Nat Rev Drug Discov 4, 477-488, doi:nrd1751 [pii] 10.1038/nrd1751 (2005).

41. Iwai, T. et al. Core 3 synthase is down-regulated in colon carcinoma and profoundly suppresses the metastatic potential of carcinoma cells. Proc Natl Acad Sci U S A 102, 4572-4577, doi:0407983102 [pii] 10.1073/pnas.0407983102 (2005).
42. Sasisekharan, R., Shriver, Z., Venkataraman, G. & Narayanasami, U. Roles of heparan-sulphate glycosaminoglycans in cancer. Nat Rev Cancer 2, 521-528, doi:10.1038/nrc842 nrc842 [pii] (2002).
43. Crocker, P. R. Siglecs in innate immunity. Curr Opin Pharmacol 5, 431-437, doi:S1471-4892(05)00072-X [pii] 10.1016/j.coph.2005.03.003 (2005).
44. Guo, Y., Mascareno, E. & Siddiqui, M. A. Distinct components of Janus kinase/signal transducer and activator of transcription signaling pathway mediate the regulation of systemic and tissue localized renin-angiotensin system. Mol Endocrinol 18, 1033-1041, doi:10.1210/me.2003-0231 me.2003-0231 [pii] (2004).
45. Rudd, P. M., Elliott, T., Cresswell, P., Wilson, I. A. & Dwek, R. A. Glycosylation and the immune system. Science 291, 2370-2376 (2001).
46. Rudd, P. M., Wormald, M. R. & Dwek, R. A. Sugar-mediated ligand-receptor interactions in the immune system. Trends Biotechnol 22, 524-530, doi:10.1016/j.tibtech.2004.07.012 S0167-7799(04)00205-7 [pii] (2004).
47. Kinjo, Y. et al. Recognition of bacterial glycosphingolipids by natural killer T cells. Nature 434, 520-525, doi:nature03407 [pii] 10.1038/nature03407 (2005).

48. Schachter, H. Congenital disorders involving defective N-glycosylation of proteins. Cellular and molecular life sciences : CMLS 58, 1085-1104 (2001).

49. Dennis, J. W., Granovsky, M. & Warren, C. E. Glycoprotein glycosylation and cancer progression. Biochim Biophys Acta 1473, 21-34 (1999).

50. Delves, P. J. The role of glycosylation in autoimmune disease. Autoimmunity 27, 239-253 (1998).

51. Gleeson, P. A. Glycoconjugates in autoimmunity. Biochim Biophys Acta 1197, 237-255 (1994).

52. Chui, D. et al. Genetic remodeling of protein glycosylation in vivo induces autoimmune disease. Proc Natl Acad Sci U S A 98, 1142-1147, doi:10.1073/pnas.98.3.1142 (2001).

53. Abbott, Karen Glycomic analysis of ovarian cancer: Past, present, and future. Cancer Biomarkers, 8(5): 1875-8592, (2011)

54. Goldberg, D.; Bern, M.; North, S. J.; Haslam, S. M. and Dell, A. Bioinformatics, 25(3): 365–371, (2009)

55. Whitehouse, C. M., Dreyer, R. N., Yamashita, M. & Fenn, J. B. Electrospray interface for liquid chromatographs and mass spectrometers. Anal Chem 57, 675-679 (1985).

56. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64-71 (1989).

57. Gaskell, S. J. Electrospray: Principles and Practice. Journal of Mass Spectrometry 32, 677-688, doi:10.1002/(sici)1096-9888(199707)32:7<677::aidjms536> 3.0.co;2-g (1997).

58. Dole, M. Molecular Beams of Macroions. J. Chem. Phys. 49, 2240 (1968).

59. MEHLIS et al. Liquid chromatography/mass spectrometry of peptides of biological samples. Vol. 352 (1997).

60. Tanaka, K. et al. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 2, 151-153, doi:10.1002/rcm.1290020802 (1988).

61. Beavis, R. C., Chait, B. T. & Fales, H. M. Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. Rapid Communications in Mass Spectrometry 3, 432-435, doi:10.1002/rcm.1290031207 (1989).

62. Beavis, R. C. & Chait, B. T. Rapid, sensitive analysis of protein mixtures by mass spectrometry. Proc Natl Acad Sci U S A 87, 6873-6877 (1990).

63. W. Paul & H. Steinwedel; Zeitschrift für Naturforschung, 8A;, p448. (1953)

64. W. Paul; Agewandte Chemie - International Edition, 29; p739. (1990)

65. G. C. Stafford et al.; International Journal of Mass Spectrometry and Ion Processes, 60; 1984, p85 and Analytical Chemistry, 59; p1677. (1987)

66. W.C. Wiley & I.H. MacLaren; The Review of Scientific Instruments, 26; 1955; p1150.

67. R.J. Cotter; Analytical Chemistry, 64; p1027A. (1992)

68. Mamyrin et al.; Soviet Physics - JETP, 37; 1973, p45 and International Journal of Mass Spectrometry and Ion Processes, 131; p1. (1994)

69. A tutorial review of TOF-MS was published in 1995: M. Guilhaus; Journal of Mass Spectrometry, 30; p1519. (1995)

70. J.A. Hipple et al.; Physical Review, 76; (1949), p1877 and Physical Review; 82; p697. (1951)

71. J.W. Cooley and J.W. Tukey; Mathematics of Computation, 19; p297. (1965)

72. M.B. Comisarow and A.G. Marshall; Chemical Physics Letters, 25; (1974), p282 and Journal of Chemical Physics, 62; (1975), p293 and Journal of Chemical Physics, 64; (1976), p110.

73. A.G. Marshall Accounts of Chemical Research, 18; 1985, p316 and Accounts of Chemical Research, 29; p308. (1996)

74. Thomas, L. H. "The Paths of Ions in the Cyclotron I. Orbits in the Magnetic Field". Physical Review 54 (8): 580–588. (1938).

75. Campbell, M., et al. Development of a pixel readout chip compatible with large area coverage, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 342, 52-58. (1994)

76. Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. Fourier transform ion cyclotron resonance mass spectrometry: A primer, Mass Spectrometry Reviews, 17, 1-35. (1998)

77. Shi, S.D.H., et al. Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry, International Journal of Mass Spectrometry, 195–196, 591-598. (2000)

78. Sleno, L., Volmer, D. and Marshall, A. Assigning product ions from complex MS/MS spectra: The importance of mass uncertainty and resolving power, Journal of The American Society for Mass Spectrometry, 16, 183-198. (2005)

79. Bossio, R.E. and Marshall, A.G. Baseline Resolution of Isobaric Phosphorylated and Sulfated Peptides and Nucleotides by Electrospray Ionization FTICR MS: Another Step toward Mass Spectrometry-Based Proteomics, Analytical Chemistry, 74, 1674-1679. (2002)

80. He, F., Hendrickson, C.L. and Marshall, A.G. Baseline Mass Resolution of Peptide Isobars: A Record for Molecular Mass Resolution, Analytical Chemistry, 73, 647-650. (2000)

81. Paul, W. Electromagnetic traps for charged and neutral particles. Reviews of Modern Physics 62, 531-540 (1990).

82. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. Mass Spectrometry Reviews 24, 1-29, doi:10.1002/mas.20004 (2005).

83. March, R. E. An Introduction to Quadrupole Ion Trap Mass Spectrometry. Journal of Mass Spectrometry 32, 351-369, doi:10.1002/(sici)1096-9888(199704)32:4<351::aid-jms512>3.0.co;2-y (1997).

84. Hager, J. W. A new linear ion trap mass spectrometer. Rapid Communications in Mass Spectrometry 16, 512-526, doi:10.1002/rcm.607 (2002).

85. Schwartz, J. C., Senko, M. W. & Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. J Am Soc Mass Spectrom 13, 659-669, doi:10.1016/S1044-0305(02)00384-7 (2002).

86. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 72, 1156-1162 (2000).

87. Hu, Q. et al. The Orbitrap: a new mass spectrometer. Journal of mass spectrometry : JMS 40, 430-443, doi:10.1002/jms.856 (2005).

88. Scigelova, M. & Makarov, A. Orbitrap mass analyzer--overview and applications in proteomics. Proteomics 6 Suppl 2, 16-21, doi:10.1002/pmic.200600528 (2006).

89. Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron Capture Dissociation of

Multiply Charged Protein Cations. A Nonergodic Process. Journal of the American Chemical Society 120, 3265-3266, doi:10.1021/ja973478k (1998).

90. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A 101, 9528-9533, doi:10.1073/pnas.0402700101 (2004).
91. Mikesh, L. M. et al. The utility of ETD mass spectrometry in proteomic analysis. Biochim Biophys Acta 1764, 1811-1822, doi:10.1016/j.bbapap.2006.10.003 (2006).
92. Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. Mol Cell Proteomics 6, 1942-1951, doi:10.1074/mcp.M700073-MCP200 (2007).
93. McAlister, G. C. et al. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. J Proteome Res 7, 3127-3136, doi:10.1021/pr800264t (2008).
94. Shi, S. D. et al. Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry. Anal Chem 73, 19-22 (2001).
95. Kelleher, N. L. et al. Localization of labile posttranslational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid. Anal Chem 71, 4250-4253 (1999).
96. Chi, A. et al. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci U S A 104, 2193-2198, doi:10.1073/pnas.0607084104 (2007).
97. Swaney, D. L., Wenger, C. D., Thomson, J. A. & Coon, J. J. Human embryonic stem cell

phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A 106, 995-1000, doi:10.1073/pnas.0811964106 (2009).

98. Mirgorodskaya, E., Roepstorff, P. & Zubarev, R. A. Localization of Oglycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. Anal Chem 71, 4431-4436 (1999).
99. Alley, W. R., Jr., Mechref, Y. & Novotny, M. V. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. Rapid Commun Mass Spectrom 23, 161-170, doi:10.1002/rcm.3850 (2009).
100. Gunl, M., Gille, S. and Pauly, M. OLIgo Mass Profiling (OLIMP) of Extracellular Polysaccharides, J Vis Exp, e2046. (2010)
101. Aoki, K. et al. Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo. J Biol Chem 282, 9127-9142, doi:M606711200 [pii] 10.1074/jbc.M606711200 (2007).
102. Aoki, K. et al. The diversity of O-linked glycans expressed during Drosophila melanogaster development reflects stage- and tissue-specific requirements for cell signaling. J Biol Chem 283, 30385-30400, doi:M804925200 [pii] 10.1074/jbc.M804925200 (2008).
103. Apte, A. & Meitei, N. S. Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. Methods Mol Biol 600, 269-281, doi:10.1007/978-1-60761-454-8_19 (2010).
104. Ceroni, A. et al. GlycoWorkbench: a tool for the computer-assisted annotation of mass

spectra of glycans. J Proteome Res 7, 1650-1659, doi:10.1021/pr7008252 (2008).

105. M. Bantsche, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, Quantitative mass spectrometry in proteomics: a critical review." Analytical and bioanalytical chemistry, vol. 389, no. 4, pp. 1017-1031, (2007).

106. R. Orlando, Quantitative Glycomics," in Functional Glycomics, J. Li, Ed. Humana Press, pp. 31-49. (2010)

107. Turnbull, J.E. and Field, R.A. Emerging glycomics technologies, Nat Chem Biol, 3, 74-77. (2007)

108. Goldberg, D., et al. Glycan family analysis for deducing N-glycan topology from single MS, Bioinformatics (Oxford, England), 25, 365-371, (2009)

109. Ceroni, A., et al. Informatics Tools for Glycomics: Assisted Interpretation and Annotation of Mass Spectra Glycoscience. In Fraser-Reid, B.O., Tatsuta, K. and Thiem, J. (eds). Springer Berlin Heidelberg, pp. 2219-2240. (2008)

110. Goldberg, D., et al. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra, PROTEOMICS, 5, 865-875. (2005)

111. Goldberg, D., et al. Automatic Determination of O-Glycan Structure from Fragmentation Spectra, Journal of Proteome Research, 5, 1429-1434. (2006)

112. Haslam, S.M., North, S.J. and Dell, A. Mass spectrometric analysis of N- and O-glycosylation of tissues and cells, Current Opinion in Structural Biology, 16, 584-591. (2006)

113. Ceroni, A., Dell, A. and Haslam, S. The GlycanBuilder: a fast, intuitive and flexible software

tool for building and displaying glycan structures, Source Code for Biology and Medicine, 2, 3. (2007)

114. Maass, K., et al. “Glyco-peakfinder” – de novo composition analysis of glycoconjugates, PROTEOMICS, 7, 4435-4444. (2007)

115. Ranzinger, R., et al. GlycomeDB - integration of open-access carbohydrate structure databases, BMC bioinformatics, 9, 384. (2008)

116. Joshi, H.J., et al. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data, PROTEOMICS, 4, 1650-1664. (2004)

117. Artimo, P., et al. ExPASy: SIB bioinformatics resource portal, Nucleic Acids Research, 40, W597-W603. (2012)

118. GlycoMod - A software Tool for Determining Glycosylation Compositions from Mass Spectrometric Data Proteomics 1:340-349 (2001).

119. Lohmann, K.K. and von der Lieth, C.-W. GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates, PROTEOMICS, 3, 2028-2035. (2003)

120. Gaucher, S.P., et al. Mass Spectral Characterization of Lipooligosaccharides from Haemophilus influenzae 2019†, Biochemistry, 39, 12406-12414. (2000)

121. Winston, W. L. Introduction to Mathematical Programming: Applications and Algorithms; Duxbury Press: Belmont, CA, (1995).

122. Lapadula, A.J., et al. Congruent Strategies for Carbohydrate Sequencing. 3. OSCAR: An Algorithm for Assigning Oligosaccharide Topology from MSn Data, Analytical Chemistry,

77, 6271-6279. (2005)

123. Ethier, M., et al. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry, Rapid Communications in Mass Spectrometry, 17, 2713-2720. (2003)

124. Tang, H., Mechref, Y. and Novotny, M.V. Automated interpretation of MS/MS spectra of oligosaccharides, Bioinformatics (Oxford, England), 21, i431-i439.(2009)

125. Horn, D., Zubarev, R. and McLafferty, F. Automated reduction and interpretation of, Journal of The American Society for Mass Spectrometry, 11, 320-332. (2000)

126. von der Lieth, C.-W., Lütteke, T. and Frank, M. The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra, Biochimica et Biophysica Acta (BBA) - General Subjects, 1760, 568-577. (2006)

127. Pedrioli, P.G.A., et al. A common open representation of mass spectrometry data and its application to proteomics research, Nat Biotech, 22, 1459-1466. (2004)

128. Ranzinger, R., Maaß, K. and Lütteke, T. Bioinformatics Databases and Applications Available for Glycobiology and Glycomics Functional and Structural Proteomics of Glycoproteins. In Owens, R. and Nettleship, J. (eds). Springer Netherlands, pp. 59-90. (2011)

129. J. Banks, J. Carson, B. Nelson, D. Nicol. Discrete-Event System Simulation. Prentice Hall. p. 3. (2001)

130. M. Better, F. Glover, G. Kochenberger, and H. Wang, Simulation Optimization: Applications in Risk Management," International Journal of Information Technology & Decision Making, vol. 7, no. 4, pp. 571-581, (2008).

131. S. Andradottir, \A Review of Simulation Optimization Techniques," in Proceedings of the 1998 Winter Simulation Conference, D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc., pp. 151-158. (1998)

132. Simulation Optimization, ser. Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice. John Wiley & Sons, Inc., ch. 9, pp. 307-333. (1998)

133. J. R. Swisher, P. D. Hyden, S. H. Jacobson, and L. W. Schruben, A Survey of Simulation Optimization Techniques and Procedures," in Proceedings of the 2000 Winter Simulation Conference, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, Eds.Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc., pp. 119-128. (2000)

134. Nelder, John A.; R. Mead. "A simplex method for function minimization". Computer Journal 7: 308–313. (1965)

135. H. H. Rosenbrock, "An Automatic Method for Finding the Greatest or Least Value of a Function", The Computer Journal 3(3): 175-184, (1960)

136. Hooke, R.; Jeeves, T.A. ""Direct search" solution of numerical and statistical problems". Journal of the Association for Computing Machinery (ACM) 8 (2): 212–229. (1961).

137. Hestenes, Magnus R.; Stiefel, Eduard (December 1952). "Methods of Conjugate Gradients for Solving Linear Systems" . Journal of Research of the National Bureau of Standards 49 (6).

138. Straeter, T. A.. "On the Extension of the Davidon-Broyden Class of Rank One, Quasi-Newton Minimization Methods to an Infinite Dimensional Hilbert Space with

Applications to Optimal Control Problems". NASA Technical Reports Server. NASA. (2011).

139. John Nelder and Saša Singer (2009) Nelder-Mead algorithm. Scholarpedia, 4(7):2928

140. Tjalling J. Ypma, Historical development of the Newton-Raphson method, SIAM Review 37 (4), 531–551, (1995). doi:10.1137/1037125

141. P. Deuflhard, Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms. Springer Series in Computational Mathematics, Vol. 35. Springer, Berlin, (2004). ISBN 3-540-21099-7.

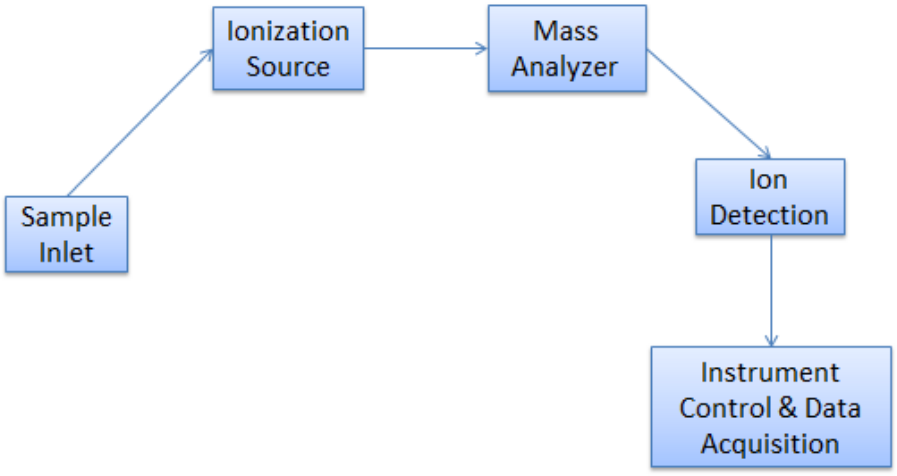


Figure 1.1. Overview of Mass Spectrometer

CHAPTER 2

A SEMI-AUTOMATIC COMPUTATIONAL FRAMEWORK FOR QUANTITATIVE GLYCOMICS

Chi Xie, Meng Fang, Lance Wells, William S. York

To be submitted to *Journal of Proteome Research*.

ABSTRACT

Although mass spectrometry (MS) has become the premier tool for quantitative glycomics, very little software is available to facilitate high-throughput quantitative glycomics analysis by MS. Several effective software tools (e.g. *GlycoWorkbench*[1], *Cartoonist*[2,3]) are available for the annotation of mass spectral features with glycan structures. Development of such tools faces many challenges, many of which arise from the complexity of glycan samples and the high probability that several glycans in a sample will have the same mass. Furthermore, new MS methods to identify and quantify individual glycans are continuously being developed, and software for processing the resulting data is always needed. We have developed a modular glycomics data processing environment and implemented some of its key components, which support semantically annotated data-exchange formats and workflow engines, enabling the development of interactive modules that perform well-defined data processing tasks. A workflow comprised of modules for data conversion, scaling, extraction, and quantification of rolling-trapping MS data was developed and implemented in this environment. This workflow was applied to experimental data, revealing sources of variation between replicates.

INTRODUCTION

Varki *et al.*[4,5] define glycomics as "profiling of the diverse glycan repertoires of cells, tissues and organisms under specific conditions". The development and application of analytical and biochemical tools for glycomics confronts several unique challenges. Due to their complex branched structures and non-template directed synthesis, glycans are more difficult to biochemically and functionally characterize than nucleic acids and proteins, which are linear molecules synthesized via template-directed mechanisms [6,7,8].

Mass spectrometry (MS) based glycomics techniques[9] are most often used to analyze free oligosaccharides that are chemically or enzymically released from glycoconjugates (e.g., glycoproteins, proteoglycans and glycolipids). However, improved MS methods for analyzing intact glycoconjugates are continuously being developed. Matrix-assisted laser desorption-ionization (MALDI)[10] and electrospray ionization (ESI)[11] are among the most commonly applied ionization methods, which can be combined with a broad range of molecular dissociation and detection methods for glycomics analysis[12]. MS may be used as a stand-alone technique, or coupled with separation methods such as high performance liquid chromatography (HPLC)[13] or gas chromatography (GC)[14].

Protein-linked *N*-glycans and *O*-glycans are typically released by enzymatic and chemical methods[15,16]. MS is particularly advantageous for the analysis of complex glycan mixtures containing oligosaccharides that cannot be identified by HPLC alone due to the lack of authentic molecular standards. MS methods that provide accurate molecular mass values

can afford direct information regarding the glycosyl residue composition of each glycan in the sample. However, several sets of isomeric glycans, each characterized by a distinct glycosyl composition and molecular mass, are often present in a sample being analyzed. The structure and abundance of each member of such an isomeric set cannot be determined by one-dimensional mass profiling techniques. More sophisticated methods such as multi-stage tandem MS (i.e., MS^n)[17,18] or combinations of MS with high-resolution separation techniques (e.g., HPLC-MS) are required to resolve the isomeric glycans in such samples.

Although many MS based glycomics tools have been developed from different perspectives, no single tool or platform can fit all the needs for glycomics research. Thus, the integration of glycomics tools to satisfy the needs for a particular glycomics project or research group is a significant challenge. Currently available MS data processing tools can accomplish many different tasks, including the following:

(1) Convert the raw data file to mzXML, a well-accepted data format for MS data exchange.

(2) Open an mzXML file for viewing and convert the data it contains into list of peaks.

(3) Detect peaks from profile-mode data file.

(4) Deisotope MS data to transform isotopic ion clusters into single entities, each represented by a single *m/z* value and a single abundance value.

Currently existing software does not include a platform or framework (analogous to the Trans-Proteomic Pipeline (TPP) software for proteomics[19]) that allows users to "personalize"

processing of glycomics data according to the input file format and the research aims. Such software, which requires the ability to select and orchestrate the appropriate functional modules designed to accomplish specific aims in the glycomics domain, are required for automatic, high throughput analysis of glycomics data.

To develop such a generic workflow system for glycomics analysis, it is first necessary to have a standard and well-accepted MS data file format. However, various mass spectrometer manufacturers have developed different proprietary formats for MS data (Table 1). The differently formatted files are very similar in information content and an integrated data processing system for mass spectra should be able to process this information irrespective of its source and original format. A common data format should facilitate the import and parsing of the MS data file as well as the exchange of experimental data sets[20,21]. The most commonly used of the currently available MS data formats is mzXML and several software applications or tools are available to convert files in other data formats to mzXML[22,23]. These include open-source parsers that can read and write files in this format (Table 2.1). We have implemented ReAdW[24] developed by Institute of System Biology to convert Thermo .raw data files to mzXML format.

The second requirement for a generic glycomics data processing system is flexibility, which can be achieved by a modular design wherein the functionality of each module is limited to a well-defined task rather than a combination of several tasks[28]. The third requirement is an intuitive graphic user interface (GUI) that facilitates the use by biologists

and biochemists as well as computer scientists[29]. This GUI should enable the user to choose appropriate functional modules, each with its own independent interface and discrete functionality, so as to form a "personalized" workflow to fit the needs of the glycomics analysis at hand. Clear documentation and user guides are necessary to make the software accessible to scientists who do not have an extensive background in computer science.

The long term goal of our research is to provide a reliable, modular semi-automatic workflow environment that is applicable to diverse MS methods for quantitative glycomics. Key features envisioned for this environment include common data exchange formats, a workflow engine to orchestrate data processing modules, an application programming interface for developing and deploying these modules as stand-alone software applications or as Web services and support for semantic annotation of the processed data. Our initial (prototype) implementation of this environment has been evaluated using the workflow described in the next paragraphs as a general use case.

Extensive analysis of mass spectral glycomics data has been performed to identify the components of sets of isomeric structures that cannot be distinguished by mass profiling alone. We have coined the term "isobaric set" to describe a set of glycans that have the same mass and share structural features because they share a common metabolic origin (i.e., are synthesized by the same cell or tissue) (Figure 2.1). An isobaric set is thus similar to a "cartoon"[30] representation that is frequently used to annotate glycan mass profiling data. The rigorous identification and quantification of the individual components of an isobaric set

for a particular biological sample requires considerable effort and extensive application of sophisticated analytical methods such as tandem mass spectrometry. However, the identities of the structures present in an isobaric set can be hypothesized based on previous analysis of similar samples (e.g., those obtained from the same tissue of different individual animals of the same species) and the assumption that the isobaric members of each set are synthesized using common biosynthetic mechanisms. Thus, quantification of the *total abundance* of all the components of an isobaric set by mass profiling has biological significance in spite of the fact that it does not establish the abundance of each and every glycan structure that is present in the set[31]. The value of the mass profiling analysis is enhanced when the structures included in each isobaric set have been assigned and available as a well-defined list in which each member is annotated with information describing the basis for the structural assignment[32]. However, processing and interpretation of the data obtained by high–throughput comparative glycomics analysis of many biological samples, even when limited to mass profiling, is extremely time consuming and thus impractical to perform using manual methods[33].

One commonly implemented mass profiling glycomics experiment involves direct infusion of the sample to an electrospray ionization source and collection of a full scan spectrum[34,35]. The ion abundance for each isomeric set of ions is determined, usually based on peak lists generated by deisotoping algorithms such as the Horn transformation [36]. This type of analysis is prone to several sources of error, including failure to correctly deconvolute

the isotopic distribution for an isomeric set of ions and failure to adequately account for the *m/z*-dependent variation in ionization efficiency. The first type of error can occur when all the members of an isomeric set are present in low abundance, resulting in low signal to noise and poor ion statistics. Signal to noise ratios can be improved by using an ion-trap instrument and filling the trap with ions within a narrow *m/z* range. A spectrum comprising the full *m/z* range can be generated by stitching several such scans together. However, the actual abundance of each ion in a given scan depends not only on the relative abundance of glycans giving rise to ions within the selected *m/z* range but also on the finite capacity of the ion trap. Therefore, direct comparison of the ion abundances in one scan to ion abundances in another scan does not necessarily provide a quantitative estimation of the relative abundances of glycans in the sample. One effective method for the abundance normalization consists of dividing the ion abundance in each scan by the length of time that the trap is collecting ions so that the normalized abundance corresponds to the ion current (ions/sec) for each species.

$N_i = t * C_i$

where N_i is the number of ions of type i in the trap when full,

C_i is the ion current due to ion i (ions of type i entering the trap per millisecond)

t is the time (milliseconds) the trap is collecting ions.

The ion current C_i is proportional to the abundance A_i of analyte i in the sample.

Then, $A_i = k * C_i = k * N_i / t$,

where k is the constant of proportionality. N_i is measured but A_i is reported in the spectrum.

An alternative approach to this problem is to record a set of spectra with overlapping *m/z* ranges and normalize the ion abundance in each such that it is the same in the overlapping region of each adjacent pair of spectra (Figure 2.2). For example, a collection of spectra with a width of 50 *m/z* are recorded, incrementing the low mass boundary by 25 *m/z* for each successive scan. The ion abundances for each scan are multiplied by a scaling factor that is calculated by analysis of the overlapping data segments, as described in the section Modules and Algorithm section.

To test our glycomics data processing environment, we have developed a workflow that implements several glycomics data processing modules to quantify the total glycan abundance for each isomeric set that is represented in a mass spectral data file generated by rolling ion-trapping methods[37]. As input, this software takes this data file along with a list of previously identified isomeric sets that are defined in an XML file. (These sets are most useful if each takes the form of a rigorously annotated isobolic set, but this is not required for quantification of the isomeric set.) This XML file specifies the glycosyl residue composition of each set along with a pointer (URI) to a database entry or ontology instance that provides, for example, information regarding the experimental basis for the structural assignment of each glycan in the set along with other meta-information (e.g., originator, generation date, etc.) This critical information provides a semantic context (that may

include subjective structural assignments) that facilitates interpretation and evaluation of the mass profiling data. The heart of the workflow is a module for the structural annotation of each mass spectral feature (isotopic distribution pattern) extracted from an experimental data set and the quantitative comparison of this feature to spectral data simulated for a specific isomeric set expected to be present in the analyte. The workflow also includes modules to visualize these results and generate a summary file specifying the estimated abundances for each isomeric set and statistical descriptors of data quality. The workflow thus includes the following modules:

1) Abundance normalization. The ion abundances are normalized for a set of mass spectra with narrow, overlapping *m/z* ranges generated by rolling-trapping methods[38]. This step is designed to minimize errors due to differences in ion-trapping efficiency at the different *m/z* ranges. This is an optional module, as we have shown that this manipulation can actually lead to an increase in the variability of the results obtained by replicate analysis (See section Implications).

2) Spectral feature assignment and extraction. Each isomeric set defined in the input file is associated with several spectral data segments that are extracted from the experimental data. The range of *m/z* values spanned by each data segment is determined by calculating the relevant isotopic distribution pattern based on the elemental composition that is a characteristic of the isomeric set, the number and identity of the ionizing species (e.g., two Na^+ ions) and the charge (e.g., 2+).

3) MS pattern simulation and verification. The isotopic distribution pattern calculated in step 2 is used to simulate an explicit profile mode spectral region with an *m/z* range corresponding to each data segment extracted in step 2. Linear regression analysis of the simulated and experimental spectra provides a correlation coefficient that can be used to judge whether spectral features in the extracted data segment arise from the target isomeric set. This analysis also provides an estimate of the total abundance of the glycans in the isomeric set and the standard error of this estimate.

4) Quantitation. The parameters calculated by regression analysis in step 3) are collected to generate output files containing the estimated abundance of each isomeric set in the sample and statistics describing the relative error and reliability of each estimate.

5) Visualization and post-processing. A graphical interface provides a visual comparison of the experimental and simulated data along with relevant statistical parameters, giving the analyst the opportunity to reject spectral feature assignments or components of the quantitation results so they are not included in the final result.

EXPERIMENTAL METHODS

Preparation of CHO-Pro5/Lec2 Protein Powder and Delipidation.

Collected CHO-Pro5/Lec2 cell pellets were dilipidated as previously reported[39,40]. Briefly, the pellets were disrupted on ice by Dounce homogenization in ice-cold 100% methanol. To extract the lipids, chloroform and water were added to the solvent mixture to give a final ratio of chloroform/methanol/water equal to 4:8:3. The resulting mixture was incubated for 3 hours at room temperature with end-over-end agitation. By centrifugation at 3,000 rpm for 15 min at 4^{o}C, the insoluble material was extracted again and incubated for 2 more hours. After washed with cold water and cold acetone three times, the final pellets of insoluble protein were dried under a stream of nitrogen.

Glycan Release and Permethylation.

The release and permethylation of *N*- and *O*-linked glycans was essentially as previously reported[39,41,42]. Briefly, the resulting protein powders were either digested by trypsin and subjected to PNGase F treatment to release *N*-linked glycans or treated with alkaline borohydride to release *O*-linked glycans by β-elimination. Released glycans were cleaned up by Sep-Pac C-18 solid-phase extraction (Sep-Pac C18 columns) and dried before they were permethylated with methyl iodide.

MS Analysis of Permethylated Glycans.

The glycans were analyzed as previously described[39,41,42] using a hybrid linear ion trap Orbitrap mass spectrometer (LTQ-Orbitrap, Thermo^{TM}). Briefly, permethylated glycans were dissolved in 15 μL of 100% methanol followed by the addition of 35 μL of 1 mM NaOH in

50% methanol to make a total of 50 μL of solution. This solution was infused directly into the mass spectrometer using a nanospray ion source with a fused-silica emitter (360 × 75 × 30 μm, SilicaTipTM, New Objective) at a syringe flow rate of 0.4 μL/min. The capillary temperature was set to 200^{o}C and MS analysis was performed in positive ion mode. Full FTMS (Fourier Transform mass spectrometry) spectra in profile mode were collected at 400-2000 *m/z* for 30 sec with 5 microscans and 1000 maximum injection time (ms) and resolution was set to be 60,000. For fragmentation by collision-induced dissociation (CID) the centroid MS/MS spectra were acquired from 400 to 2000 *m/z* at 36% normalized collision energy, 0.25 activation Q, and 30.0 ms activation time by total ion mapping (TIM). Parent mass step size and isolation width was set at 2.0 *m/z* and 2.8 *m/z* respectively for automated MS/MS spectra with TIM scans. All glycan structures were manually interpreted based on in-house fragmentation rules and *GlycoWorkBench* software[1].

For the quantification experiment, a newly designed instrumental method called rolling trapping scans was applied to increase the signal to noise ratio of full mass spectra and present more isotopic peaks for each individual glycan structures. Automated MS/MS spectra without applying collision energy were acquired in this mode. For each MS/MS scan, the instrument trapped the ions with a width of 100 *m/z* but only scanned 50 *m/z* around the center of the trapping window to ensure the trapping efficiency. The trapping center of each scan was moved automatically by 25 *m/z* to the higher mass region compared to its previous scan so that the range from 400 to 2000 *m/z* was covered by these stepwise trapping scans.

Modules and Algorithm

Computational modules were developed to perform data-processing tasks described above. Data exchange between modules was accomplished using extensible markup language (XML)[43] to save and keep track of all the input variables, the intermediate results as well as the final results. This section describes the algorithms implemented by each module.

Abundance Normalization of Rolling-Trapping Scans

The initial step performed by the normalization module is to identify overlapping data segments for adjacent scans in a data file that has been generated by rolling-trapping mass spectrometry methods[44,45,46]. The ion abundances in the overlapping segments are equalized by multiplying one by a scaling factor calculated by linear regression of the abundance values in the two data segments. This point-by-point comparison requires mapping of every data point in one data segment with a corresponding point in the other. Since the *m/z* values for data points in the two segments do not necessarily align, the data must be transformed such that each data point in one data segments corresponds to a unique data point with the same *m/z* value in the other data segment. This is accomplished by an interpolation algorithm **(**described in Chapter 5**)** that is applied to both data segments to generate precisely matching data segments with points that are evenly spaced on the *m/z* axis.

Spectral Feature Annotation and Extraction

In the context of this software described here, a spectral feature is defined as a pattern of signals corresponding to a cluster of ions arising from a collection of molecules that have the

same elemental composition. Such an isomeric set can include one or more molecular structures. The ions in the cluster correspond to the various isotopologues that exist as a result of the presence of isotopes in the sample. An isotopologue is defined by the IUPAC Gold Book[47] as "A molecular entity that differs only in isotopic composition". Thus, two isotopologues of ethanol are $^{12}C_2H_6O$ and $^{12}C^{13}CH_6O$, which have the same elemental composition but different masses. The $^{12}C^{13}CH_6O$ isotopologue consists of two distinct isotopomers, *i.e.,* [1-^{13}C]-ethanol and [2-^{13}C]-ethanol.

A primary task in processing mass spectral data is the annotation of spectral features with the identities of specific components of the analyte. Various approaches for associating structures with spectral features have been implemented, including tandem mass spectrometry (MS^n), which provides direct structural information for ions that are selected from a survey scan. In high-throughput quantitative glycomics analysis[48], the structures of the glycans present in the analyte may be previously known or assumed to be known, and the task is then to associate each glycan or set of isomeric glycans present in the analyte with a spectral pattern and quantify the ion abundance giving rise to that pattern. Although straightforward, this task is very time consuming for the human analyst and manual data processing is not practical for high-throughput data.

One approach for identifying and assigning spectral features involves peak picking to obtain a "peak list" of spectral signals, each of which is characterized by its centroid *m/z* ratio and abundance, followed or preceded by "deisotoping", which provides another peak list

containing one member for each spectral feature (isotopic pattern) in the spectrum. The peaks in this list are associated with a structure (e.g., from a database) and quantified. Deisotoping is often performed using a Horn transformation, which compares each observed signal pattern to a theoretical pattern based on a mass-dependent average elemental composition. In the case of proteomics, the average elemental composition is based on the concept of averagine[49], a contrived amino acid whose elemental composition corresponds to the weighted average elemental composition of amino acids in the analyte. For glycomics, one can use a similar approach to define an "averagose"[50]. Various methods are then used to associate each deisotoped peak obtained by this process to a specific structure or set of isomeric structures.

Our alternative approach to quantitative glycomics is to use prior knowledge regarding the identity of glycans present in the analyte. The elemental composition of each glycan expected to be present in the analyte is readily calculated, allowing the spectral pattern for that glycan to be simulated and compared to the patterns observed in the real spectrum. This approach has several advantages, including elimination of the requirement for accurate peak picking and the elimination of errors introduced by using average elemental compositions rather than precisely defined elemental compositions. This approach is more applicable to glycomics, where one expects to observe spectral features at a relatively small number of discrete *m/z* values, compared to proteomics, where ions can be observed at virtually any *m/z* value.

We have developed feature extraction and spectral annotation modules that process an input file containing mass spectral data using information provided in another XML-encoded input file containing a list of structures and structural parameters. Each member of this list has a distinct elemental composition and can be an individual glycan or a set of isomeric glycans, depending on the complexity of the sample. Several spectral features can be associated with each elemental composition, as each molecule can form different ions with distinct charges (*z*) by association with different ionizing species (e.g., H^+, Na^+ or K^+). Using this information, the *m/z* range spanned by the isotopic distribution for each ionic species is calculated and a data segment containing the predicted spectral feature is extracted from the mzXML input file (Figure2.3). When the rolling-trapping method is used to record overlapping mass spectra, the spectral segment is taken from the scan with an *m/z* range that places the spectral feature closest to the center of the scan, avoiding edge effects. The extracted data segments are saved in an output file (XML format) along with meta data specifying the identity of the corresponding glycan or set of isomeric glycans, the charge *z* and the ionizing species for each segment. The data processing module thus performs a specific data extraction and annotation task that can be applied to diverse mass spectral data sets containing spectral features that correspond to known or predicted structures.

The *m/z* range for each data segment generated by this module is determined by identifying the significantly populated isotopologues for an ionic structure arising from a specific member of the structure list. Each structure in this list is defined by its glycosyl

composition, which, in combination with parameters defining any chemical derivatization (e.g., per-*O*-methylation), structural features of the reducing end of the glycan (e.g., conversion to an alditol), and the identity and number of ionizing species (e.g., Na^+), are used to calculate an elemental composition for each expected ion. The *m/z* ratio and the relative abundance of each isotopologue corresponding to this composition is calculated from first principles using methods defined in the JMass API[51]. A coverage parameter (usually set to 99%) limits the calculation to isotopologues that are significantly populated, drastically reducing the time required to generate the list of isotopologues. The *m/z* range of this list is used as a guide to define the *m/z* range to be extracted from the mass spectral data. XML parsers, DOM4j[52,53] and JDOM[54,55] are implemented to locate the scan whose center is closest to the monoisotopic mass for the selected ion, and the data segment is extracted from that scan (Figure 2.3).

Often data points in the mass spectral file are not evenly spaced along the *m/z* axis. For example, data points may not be explicitly described for stretches in which the abundance is zero. However, data points in the same *m/z* range in a theoretical spectrum simulated using an elemental composition may have non-zero abundances. Therefore, to facilitate valid comparisons of the extracted data segment to a simulated data segment, data points that are uniformly spaced along the *m/z* axis are explicitly included in the output file. This is performed using an interpolation algorithm described in Chapter 5.

The header of the output XML file contains data acquisition and sample identity information from MS data file, along with parameters describing the desired isotopologue coverage, charge range (e.g., +1 to +4), adduct identities, chemical derivatization and reducing-end structure. Other general information specified in the header includes the name of the original file containing mass spectral data, the name of file containing the list of glycan structures, the name of the individual who defined the structure list, and the date on which the output file was generated. As listed in Table 2.2, each data segment in the output file is annotated with the glycosyl and elemental compositions of the members of this set, the charge state, the identity and number of charge carriers, the isotopologue coverage, the base peak intensity, the base peak *m/z*, the highest and lowest value of *m/z* and the scan number from which the data segment was extracted and a pointer (e.g., URL) to a digital resource specifying the members of a set of isomeric structures assigned to a spectral feature in the segment.

COMMENTARY

Evaluation of MS Pattern Annotation and Quantification of Ion Abundance

After a spectral segment has been extracted and annotated for each item in the structure list, another module is implemented to statistically evaluate the annotation for each segment and quantify the ion abundance for each annotated spectral feature. Manual data analysis often involves selecting and annotating a specific scan containing a spectral feature corresponding to the target structure and charge state. As this process is usually based on personal experience and may involve manual calculation of *m/z* values, the results are often subjective and error-prone. Signals of some analytes may be neglected or noise may be inappropriately annotated as arising from a glycan that is not actually present in the sample. Statistical methods to evaluate the confidence in the assignment or quantification are too rarely employed when the data are processed manually.

Our approach to address these issues is to simulate the isotopic distribution for each spectral feature using an optimized set of parameters and then evaluate the quality of the annotation by comparing the extracted experimental spectral segment with the simulated data. The initial step performed by this module is to optimize parameters describing the spectral (*m/z*) calibration offset (δ) and the peak width (*pw*). This is accomplished by simulation of the monoisotopic peak, which corresponds to a single isotopologue. A small data segment containing only the monosisotopic peak is extracted from the scan and a linear regression is performed to determine the correlation of this data segment with the simulated peak. This

requires a point-by-point mapping of the two data sets. Therefore the experimental data processed by this module should already be transformed using an interpolation algorithm to generate a new data set that is uniformly spaced along the *m/z* axis. This step is implemented using the data extraction module described above. The spectral simulation uses the transformed set of *m/z* values (after taking the calibration offset δ into account) as the dependent variable to calculate theoretical abundance data. The Pearson correlation function ρ is used as the objective function for the optimization, which is implemented using the Nelder-Mead algorithm[56], also known as simplex optimization algorithm.

The resulting optimized values of δ and *pw* are then used to simulate the entire spectral segment containing the theoretical isotope distribution pattern for the target structure. Linear regression is again used to correlate the extracted spectral data with the corresponding simulated data. In this case, the Pearson correlation coefficient ρ provides a statistical measure of the quality of the assignment and the slope *m* of the regression curve provides an estimate of the ion abundance for the spectral feature (Figure 2.3). The error of the slope ε is also calculated, providing an estimate of the accuracy of the quantification.

For all simulations, each peak is assumed to have Gaussian[57] line shape. The monoisotopic peak signal distribution is calculated as a vector **S**, whose elements s_i are

$$s_i = \exp\left(\frac{-\left(m_{mono} - x_i\right)^2}{2\sigma^2}\right)$$

where,m_{mono} is the monoisotopic mass, x_i is *m/z* for data point *i*, and σ describes the width of the Gaussian distribution. The parameter σ is related to *pw*, defined as the full width at half maximum (FWHM). That is $\sigma = 0.4247\ pw$[58].

The entire spectral pattern is simulated based on the set of isotopologues whose abundances a_j and masses m_j are calculated using the jMass API. In this case the simulated data consists of a normalized sum of Gaussian distributions to generate the vector **S**.

$$s_i = \frac{1}{\sigma\sqrt{2\pi}} \sum_j a_j \exp\left(\frac{-\left(m_j - x_i\right)^2}{2\sigma^2} \right)$$

Here, a_j and m_j are the abundance and mass of isotopologue *j*, respectively.

Glycan Quantification

After the predefined glycan structures have been annotated within the MS data file and the corresponding data segments have been extracted, the glycan quantification module is implemented to quantify each structure in the list. The results provide a basis for understanding the consequences of disease, developmental stage, tissue or cell selection, genetic diversity, *etc*. on the abundance of these structures. It is critical to provide statistics that reveal the accuracy and reproducibility of this quantification, which is another important implementation within this module.

In our approach, the first implemented algorithm is to make the comparison between the experimental mass spectral segment annotated to each glycan structure in the list with the simulated mass spectrum pattern from the module described above and some of the statistical parameters from the above modules will be used as initial estimation for the abundance or

amount for the isobaric set. To do this, the module implements a linear regression[59,60,61] analysis correlating the calibrated experimental data to the simulated data (Figure 2.4). As shown in Figure 2.4, the slope of the linear correlation line corresponds to the area of a defined spectral feature in the experimental data segment, relative to its simulation. The mathematical form of the Gaussian simulation (above) ensures that the area under the simulated peaks is normalized to 1.0. Additional data-processing methods are implemented in subsequent steps to account for noise or contaminant signals in the spectrum. These include the *clean* method, to reduce noise and contaminant signals, and the *purge* method, to quantify signals with poor signal to noise in a way that is analogous to the manual method used by some analysts. (See section Results and Discussions for details)

The short-term objective associated of our glycomics data processing framework is to do a mass profiling study[62,63] to detect and quantify differences in glycosylation for different cell lines. Solving this problem requires the collection of the linear regression analysis parameters and the estimated value of abundance for each isobaric set, along with the associated statistics, which describe the reliability and confidence for each of the estimations. This information is saved as an output file that is generated by this module. As all the intermediate analysis results and parameters may have relevance for interpreting the data, all the information in the file generated by the previous module is kept and saved in the resulting XML file.

Visualization and post-processing

Previously described modules perform spectrum annotation and quantification of glycans to generate intermediate results as well as the final results. However, the results are saved in XML format, which is a not a human-readable format. For example, the extracted experimental segments are saved as base64 format[64] to save space. Therefore, we developed a module that provides graphical representation of these results, allowing the user to visually compare experimental and simulated spectra while viewing statistical measures of their correlation. This allows the user to determine whether each annotation and quantification result should be included in the final output.

This module provides an interface showing the list of structures along with the statistics from the quantification module. Profiling results are calculated for each ionic species (with its distinct glycan, charge carrier and overall charge). These are combined to generate results for each glycan structure. These two data sets are presented and manipulated by two separate interfaces and recorded as two separate *csv* files.

The final result can be viewed as a pseudo-spectrum in which the abundance of each glycan structure is represented as a vertical bar located at the horizontal position corresponding to its mass (Figure 2.5, Table 2.3). The abundance data for individual sets of isobaric ions at each charge state are selected for inclusion in this representation depending on whether the data results in a value of ρ that is above a specified threshold value (e.g., 0.9). A complete list of abundances and statistics, including ρ and the relative error is also available in tabular form for the original data, the *cleaned* data and the *purged* data.

Results and Discussion

Evaluation of workflow for processing rolling-trapping scan data.

To test the efficiency and accuracy of our framework, we use the rolling trapping MS data obtained by analysis of *O*-linked glycans from Chinese Hamster Overary (CHO)-Pro5 and CHO-Lec2[65,66,67] cells as the case to study. The *O*-glycans were extracted and per-*O*-methylated according to the protocol described in the Methods section. The prepared samples were then analyzed using a Thermo orbitrap spectrometer to generate rolling-trapping spectral data. The data files from these two CHO lines, Pro5 and Lec2 were processed using the modular workflow framework. The first module converted the data from .raw format into mzXML format. The abundance values in each data file were scaled by comparing overlapping (*m/z*) sections of scans within each output file. (This last step was optional and omitted from some data processing runs.) Data segments corresponding to each possible charge state of each glycan structure in the list were then extracted from the mzXML data file using the glycan structure annotation module, which selects and extracts an appropriate *m/z* range to extract for each glycan structure in the list. This selection is based on a calculating the *m/z* values for the most abundant isotopologues for each glycan. In one example (the per-*O*-methylated oligoglycosyl form of $Hex_1HexNAc_1$), the range of the isotopologue distribution is from *m/z* 534.289 to *m/z* 538.3, which was used to select scan #3 from the mzXML file, extract the data segment spanning the range 533.1892 to 538.3933 from this scan (Figure 2.3) and save the segment as a record in the intermediate result file.

Signals in this data segment correspond to the ionized glycan ($Hex_1HexNAc_1$), together with the noise and ions that arise from contaminants in the sample.

Glycan abundances are estimated by applying the glycan quantification module to the extracted data segments. This module includes the so-called *clean* and *purge* methods to generate data from which the noise and contaminant peak areas have been excluded. Initially, glycan quantification was carried out by using the data segments extracted from data files that were normalized by comparing the overlapping regions of adjacent scans. A summary of these results is presented in Figure 3.8 for the raw data, and the results of the *clean* and *purge* data processing methods.

In the example, the extracted segment corresponding to ($Hex_1HexNAc_1$) from CHO-Pro5 cells was processed with the *clean* method, which is designed to remove noise and contaminating signals by selecting only those data points in the experimental data set that correspond to data points with simulated abundance values that are above a specified threshold, as shown in Figure 2.4B-1 and 2.4B-2. This corresponds to removal of points that lie close to the vertical axis in the plot of the original data (Figure 2.4A-2). Applying the *clean* method increased the value of ρ from 0.84 to 0.93, indicating that it is an effective method to remove the noise as well as contaminating ions from consideration, as the result contains only those signals that co-localize with simulated signals.

Based on the results processed with the *clean* method, the extracted segments for ($Hex_1HexNAc_1$) from CHO-Pro5/Lec2 was further processed using the *purge* method, which operates on *cleaned* data and purges signals from the simulated spectrum where the

experimental signal intensity is below a specified threshold. This corresponds to removing data points near the horizontal axis in Figure 2-4B-1 and mimics a common manual data processing method to obtain quantitative information from ion clusters with signal to noise ratios that are so low that some of the isotopologue peaks are missing in the experimental data. (Details as shown in Chapter 5) The *purge* method slightly improved the value of ρ to 0.94, which indicates good agreement between the modified experimental and simulated spectra. (Figure 2.4.C) The *purge* method has little or no effect on the results for high-abundance signals, but can have significant effects in the results for ion clusters with low abundance, as shown in Figure 2.4.C-1**.** Thus, although the purge method can be useful in increasing the number of glycan structures in the final output, it is likely to result in artifacts that increase the reported statistical confidence for low abundance structures, which can be assigned misleading correlation coefficients and relative errors. Thus, signals whose correlation coefficients increase significantly when the *purge* method is applied can be listed in the final results (based on the improved correlation score) but the quantification of these signals should not be considered to be trustworthy. This situation occurs most often for low abundance glycans, which should be annotated as "probably present" rather than with an untrustworthy numerical abundance estimate. These methods provide the careful analyst with meaningful estimations of glycan abundances in the sample.

The reproducibility of the overall approach (including sample analysis in the laboratory and data processing) was evaluated by comparing three replicate data sets for glycans isolated from the CHO-Pro5 cells **(**Figure 2.6**)**. Our initial results were not highly

reproducible, with a large variance from one replicate to the next (Table 2.4, Table 2.5). The source of this variance was identified as variability in the scaling of the rolling trapping data. The most consistent results for high-abundance structures was obtained using full-scan data, but these data did not give satisfactory results for low abundance structures. Rolling trapping data that was normalized using the default algorithms provided by Thermo Scientific provided reasonably reproducible results for both high and low abundance structures. Scaling by comparing overlapping segments of rolling trapping scans was improved by using a modified scaling algorithm that removes data points that are outliers in the linear regression of the overlapping data (Figure 2.6).

Implementing the modified normalization module to the CHO-Pro5 data showed that, as more outliers are rejected, the reproducibility increases and the results approach those obtained without application of the normalization module (Figure 2.6). However, all data was (by default) processed with the Thermo data normalization method before applying our normalization method, which consistently lead to a decrease in the reproducibility of the output. This is likely to occur because errors occurring when normalization factors are calculated by evaluating overlapping regions of adjacent scans are combined and propagated to other scans while normalization factors based on the ion-trap filling time are independent for each scan. These results indicate that the scaling algorithm, even when improved by rejecting outliers, should not be used for unless there is no alternative effective method for scaling the individual scans.

Statistical analyses were carried out and the linear regression model was applied to calculate the abundances and standard errors of the abundance estimations for three runs of CHO-Pro5 samples. Based on statistical analysis for three sample runs performed using the Thermo Orbitrap without application of scaling methods, good reproducibility was obtained, as reflected in low values for the relative standard errors for each glycan structure (Figure 2.6). However, it was not possible to determine the absolute abundances from different runs because the samples did not contain a valid internal standard.

Based on the list of glycan structures as one of the input files, signals corresponding to those structures with high abundances can be identified and quantified with ρ values approaching 1.0 and relative small standard error (Figure 2.7). Our workflow can also identify and quantify structures with lower abundances, although these usually result in data with lower ρ values and higher standard errors as a result of noise and contaminant signals, which have a relatively high effect on the statistics for these structures (Figure 2.8).

To test the sensitivity of our framework, the relative standard error (RSE)[68] of the slope of the regression for the correlation of annotated experimental data segments with the simulated spectrum was correlated to the percent of data segments that are acceptable based on the criterion that they have a relative standard error that is below a specified threshold value (Figure 2.9). As expected, increasing the error threshold increases the number of acceptable data segments. Compared with full-san data set, rolling-trapping data sets have a higher fraction of acceptable data sets at the same threshold values, indicating that this data

acquisition method has better sensitivity for the identification and quantification of glycan structures.

Implications

The development of sensitive methods of mass spectrometry, e.g. rolling-trapping method, has allowed the glycan complement of cells to be examined in detail. Annotation of spectral features is facilitated by previously acquired knowledge of the metabolic pathways involved in glycosylation. Such knowledge sometimes allows variations in glycan abundance to be correlated to variations in the flux through specific pathways that occur as cells develop or are genetically altered. We performed a test-case analysis in order to assess the suitability of our data processing platform to provide results that are relevant in this context. Application of the data processing software allowed us to identify and compare the *O*-glycosylation characteristics of Chinese hamster ovary parent (CHO-Pro5) cells and cells of the glycosylation mutant, CHO-Lec2. Rolling-trapping data sets (processed using the default Thermo normalization algorithm for abundance scaling) were converted to mzXML format and annotated data segments were extracted. For the quantitative analysis, the threshold of value of ρ set to 0.9 to reject unreliable results for each possible charge state and slopes obtained by linear regression analysis were used to calculate the prevalence or each isobaric glycan set (Table 2.6).

Hypothesis

CHO glycosylation mutants isolated on the basis of their lectin resistance have been particularly useful for glycosylation engineering of recombinant glycoproteins. Although

several lectin-resistance CHO mutants have been previously characterized biochemically and genetically, their *O*-linked glycans have not been fully characteized[69].

A previous study showed that, compared with the wild type Pro5 cells, Lec2 cells produce *N*-linked glycans with a decreased amount of sialic acid[70]. The most abundant *N*-linked glycans produced by the Lec2 mutant are asialo, core-fucosylated structures. As this effect is due to deficiency of Golgi-localized CMP-sialic acid (SA) transporter activity, the profile of *O*-linked glycan structures in the Lec2 cell line is likely to be affected in a similar way (i.e., a reduction in sialylation)[71](Figure2.10).

The following hypotheses can be made based on these observations:

i) *O*-glycans within CHO cells include *O*-fucosylated, *O*-glucosylated, or *O*-mannosylated structures[72].

ii) There is a deficiency of Golgi-localized CMP-sialic acid transporter activity in CHO-Lec2 cells compared to CHO-Pro5 cells[73].

iii) Some sialylated *O*-glycan structures may still be produced if low levels of CMP-sialic acid can find their way into the Golgi of CHO-Lec2 cells.

Although *O*-glycans from the cell lines CHO-Pro5 and CHO-Lec2 cell lines have been previously compared[71] further analysis of these cells should provide additional information regarding the relationships between the genetic makeup of the two lines and their glycomics profiles.

Testing of Hypothesis and Glycomics Profiling Study

Our profiling study of *O*-glycans for these two cell lines shows that the percentage of sialic acid containing structures is 97.7% in CHO-Pro5 and it is decreased to 14.5% in

CHO-Lec2 cells. This confirms the hypothesis that the CHO-Lec2 has a collection of sialic acid-deficient *O*-linked oligosaccharides since CHO-Lec2 cell line is deficient in CMP-sialic acid translocase activity.

In our glycomics profiling study for *O*-glycan, as shown in Figure 2.11, the prevalence for the isobaric set ($Hex_1HexNAc_1$) is increased in the mutant of CHO-Lec2 to 83% compared with the amount of 2.29% in CHO-Pro5. On the other hand, the prevalence for the glycans with the compositions ($NeuAc_1Hex_1HexNAc_1$) and ($NeuAc_2Hex_1HexNAc_1$) is decreased from 63.02% and 30.7% to 13.75% and 0 respectively, which indicates the decreased activity of CMP-sialic acid translocases and a resulting deficiency in the cells' ability to transport CMP-sialic acid from the cytosol to Golgi. Surprisingly, Lec2 cells still produce low amounts of sialylated glycan structures, indicating either that the mutation leading to the Lec2 phenotype is leaky or that an alternative mechanism exists for putting CMP sialic adic in the Golgi.

CONCLUSION

Rapid advancement of experimental techniques in MS strategies requires the emergence of new quantitative and analytical methodologies from the computational side for high-throughput MS data processing. Therefore, it is urgent to develop quantitative software tools capable of processing and interpreting MS raw data generated from these strategies. This framework supports semi-automated quantitative software for identification and annotation of spectral features generated by rolling trapping mass spectroscopy and the label free quantification of isobaric glycans. The estimation and simulation in the distribution of isobaric sets of glycans produced by different cell lines incubated in the same condition can provide key information regarding the study of diverse biological processes such as glycan metabolism and biosynthesis. This kind of information is difficult to obtain by other currently available methods or platforms.

Evaluation of this framework for the processing of MS profile data lead to the following conclusions. (i) The MS data processing software is a convenient tool for the rapid assignment of spectral features in rolling trapping mass spectral data, providing valuable quantitative information, including statistical parameters that allow the reliability of each measurement to be assessed in the context of the entire data set. (ii) Little human intervention is required for the software to process the data in a short period of time. (iii) The software provides a graphical interface that facilitates human examination of data points with questionable statistical parameters. (iv) Rolling-trapping data acquisition methods provide better ion statistics for low abundance analytes than do full scan methods. (v) Scaling of

rolling trapping mass spectra by analysis of the overlapping regions of adjacent spectra leads to scaling artifacts that become unacceptably large as they are propagated through the entire data set. (vi) Low abundance ions give rise to spectral features with poor statistics and estimated sample to sample abundance differences are not trustworthy for such analytes.

Successful implementation of the software requires compilation of a reasonable list of candidate structures (expressed as glycosyl compositions) that have been modified in various ways (e.g., by per-*O*-methylation). The modular work-flow structure supported the platform makes it flexible such that it can be adapted to rapidly changing methods for glycan analysis by introduction of modules that perform well-defined data processing tasks.

REFERENCE

1. Ceroni, A., et al. GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans†, Journal of Proteome Research, 7, 1650-1659. (2008)

2. Goldberg, D., et al. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra, PROTEOMICS, 5, 865-875. (2005)

3. Goldberg, D., et al. Automatic Determination of O-Glycan Structure from Fragmentation Spectra, Journal of Proteome Research, 5, 1429-1434. (2006)

4. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3, 97-130 (1993).

5. Cohen, M. and Varki, A. The sialome--far more than the sum of its parts, Omics : a journal of integrative biology, 14, 455-464. (2010)

6. Uematsu, R., et al. High Throughput Quantitative Glycomics and Glycoform-focused Proteomics of Murine Dermis and Epidermis, Molecular & Cellular Proteomics, 4, 1977-1989. (2005)

7. Horai, H., et al. MassBank: a public repository for sharing mass spectral data for life sciences, Journal of Mass Spectrometry, 45, 703-714. (2010)

8. Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C. & Sasisekharan, R. Glycomics: an integrated systems approach to structure-function relationships of glycans. Nat Methods 2, 817-824, doi:10.1038/nmeth807 (2005).

9. F. Forner, L. J. Foster, and S. Toppo, \Mass Spectrometry Data Analysis in the Proteomics Era," Current Bioinformatics, vol. 2, pp. 63{93(31), (2007)

10. Morelle, W., Faid, V., Chirat, F. & Michalski, J. C. Analysis of N- and O-linked glycans from glycoproteins using MALDI-TOF mass spectrometry. Methods Mol Biol 534, 5-21, doi:10.1007/978-1-59745-022-5_1 (2009).

11. Pabst, M. & Altmann, F. Glycan analysis by modern instrumental methods. Proteomics 11, 631-643, doi:10.1002/pmic.201000517 (2011).

12. Aoki-Kinoshita, K. F. An introduction to bioinformatics for glycomics research. PLoS Comput Biol 4, e1000075, doi:10.1371/journal.pcbi.1000075 (2008)

13. Snyder, L.R., Kirkland, J.J. and Glajch, J.L. Practical HPLC method development. John Wiley & Sons. (1997)

14. Grob, R.L. and Barry, E.F. Modern practice of gas chromatography. Wiley-Interscience. (2004)

15. Dell, A. & Morris, H. R. Glycoprotein structure determination by mass spectrometry. Science 291, 2351-2356 (2001).

16. Haslam, S. M., North, S. J. & Dell, A. Mass spectrometric analysis of N- and Oglycosylation of tissues and cells. Curr Opin Struct Biol 16, 584-591, doi:10.1016/j.sbi.2006.08.006 (2006).

17. M. Bantsche , M. Schirle, G. Sweetman, J. Rick, and B. Kuster, Quantitative mass spectrometry in proteomics: a critical review." Analytical and bioanalytical chemistry, vol. 389, no. 4, pp. 1017-1031, (2007)

18. R. Orlando, Quantitative Glycomics," in Functional Glycomics, J. Li, Ed. Humana Press, pp. 31-49. (2010)

19. Keller, A. and Shteynberg, D. Software Pipeline and Data Analysis for MS/MS Proteomics:

The Trans-Proteomic Pipeline Bioinformatics for Comparative Proteomics. In Wu, C.H. and Chen, C. (eds). Humana Press, pp. 169-189. (2011)

20. Pedrioli, P.G.A., et al. A common open representation of mass spectrometry data and its application to proteomics research, Nat Biotech, 22, 1459-1466. (2004)

21. Lin, S.M., et al. What is mzXML good for?, Expert Review of Proteomics, 2, 839-845. (2005)

22. Sturm, M., et al. (2008) OpenMS - an open-source software framework for mass spectrometry, BMC bioinformatics, 9, 163.

23. [37] Martens, L., et al. mzML—a Community Standard for Mass Spectrometry Data, Molecular & Cellular Proteomics, 10. (2011)

24. Falkner, J.A., Falkner, J.W. and Andrews, P.C. ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats, Bioinformatics (Oxford, England), 23, 262-263. (2007)

25. Riley, C.P., et al. The Proteome Discovery Pipeline–A Data Analysis Pipeline for Mass Spectrometry-Based Differential Proteomics Discovery, Open Proteomics Journal, 3, 8-19. (2010)

26. Tuli, L., et al. Using a spike-in experiment to evaluate analysis of LC-MS data, Proteome Science, 10, 13. (2012)

27. For a list of C-mannosylation sites see - Julenius K. (2007)

28. “NetCGlyc 1.0: Prediction of mammalian C-mannosylation sites.” Glycobiology. 17:868-76. (2010)

29. von der Lieth, C.-W., et al. EUROCarbDB: An open-access platform for glycoinformatics, Glycobiology, 21, 493-502. (2011)

30. Bouwman, B., Bruinsma, I. and van Rijn, C. The interaction between vigabatrin and R-baclofen on absence epileptic spike and wave discharges in WAG/Rij rats; an isobolic analysis, GABAergic drugs in WAG/Rij rats, 43.(2009)

31. Peltoniemi, H., et al. Novel data analysis tool for semiquantitative LC-MS-MS&lt profiling of N-glycans, Glycoconjugate Journal, 1-12. (2009)

32. Aoki, K., et al. Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo, The Journal of biological chemistry, 282, 9127-9142. (2007)

33. North, S.J., et al. Mass spectrometry in the analysis of N-linked and O-linked glycans, Current Opinion in Structural Biology, 19, 498-506. (2009)

34. Lee, W.N.P., et al. Mass isotopomer analysis: Theoretical and practical considerations, Biological Mass Spectrometry, 20, 451-458. (1991)

35. Zaia, J. Mass Spectrometry and the Emerging Field of Glycomics, Chemistry & Biology, 15, 881-892. (2008)

36. Horn, D., Zubarev, R. and McLafferty, F. Automated reduction and interpretation of, Journal of The American Society for Mass Spectrometry, 11, 320-332. (2000)

37. R. H. Perry, R. G. Cooks, and R. J. Noll, Orbitrap mass spectrometry: Instrumentation, ion motion and applications," Mass Spectrometry Reviews, vol. 27, no. 6, pp.661-699, (2008)

38. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 72, 1156-1162 (2000).

39. Orlando, R. et al. IDAWG: Metabolic incorporation of stable isotope labels for quantitative glycomics of cultured cells. J Proteome Res 8, 3816-3823, doi:10.1021/pr8010028 (2009).

40. Fang, M., Lim, J.-M. & Wells, L. Quantitative Glycomics of Cultured Cells Using Isotopic Detection of Aminosugars with Glutamine (IDAWG). (2009).

41. Aoki, K. et al. Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo. J Biol Chem 282, 9127-9142, doi:M606711200 [pii] 10.1074/jbc.M606711200 (2007).

42. Aoki, K. et al. The diversity of O-linked glycans expressed during Drosophila melanogaster development reflects stage- and tissue-specific requirements for cell signaling. J Biol Chem 283, 30385-30400, doi:M804925200 [pii] 10.1074/jbc.M804925200 (2008).

43. Arenas, M. and Libkin, L. XML data exchange: Consistency and query answering, J. ACM, 55, 1-72, (2008)

44. Van Berkel, G.J., Glish, G.L. and McLuckey, S.A. Electrospray ionization combined with ion trap mass spectrometry, Analytical Chemistry, 62, 1284-1295. (1990)

45. Le Blanc, J.C.Y., et al. Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications, PROTEOMICS, 3, 859-869. (2003)

46. McAlister, G.C., et al. A Proteomics Grade Electron Transfer Dissociation-Enabled Hybrid Linear Ion Trap-Orbitrap Mass Spectrometer, Journal of Proteome Research, 7, 3127-3136. (2008)

47. IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A.

D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook. (2006)

48. Park, S.K., et al. A quantitative analysis software tool for mass spectrometry-based proteomics, Nat Meth, 5, 319-322. (2008)

49. Yao, X., et al. Averagine-scaling analysis and fragment ion mass defect labeling in peptide mass spectrometry, Analytical Chemistry, 80, 7383-7391. (2008)

50. Kronewitter, S.R., et al. The glycolyzer: Automated glycan annotation software for high performance mass spectrometry and its application to ovarian cancer glycan biomarker discovery, PROTEOMICS, 12, 2523-2538. (2012)

51. Han, J., et al. GlycoQuant: An Automated Simulation Framework Targeting Isotopic Labeling Strategies in MS-Based Quantitative Glycomics. Technical report. (2011)

52. WANG, J. and HUANG, C. Application of Dom4j in Data Exchange, computer and modernization, 5, 98-99. (2007)

53. Wei, L. Study of Technology Applied to Database Integration Based on XML+ DOM4J [J], Computer Knowledge and Technology (Academic Exchange), 5, 006. (2007)

54. Harold, E.R. Processing XML with Java: a guide to SAX, DOM, JDOM, JAXP, and TrAX. Addison-Wesley Professional. (2003)

55. Hunter, J. JDOM makes XML easy. Sun's 2002 Worldwide Java Developer Conference. (2002)

56. Powell, Michael J. D. "On Search Directions for Minimization Algorithms". Mathematical

Programming 4: 193–201, (1973)

57. Hagen, N. and Dereniak, E.L. Gaussian profile estimation in two dimensions, Applied optics, 47, 6842-6851. (2008)

58. Weisstein, E.W. Gaussian function, From MathWorld–A Wolfram Web Resource. http://mathworld. wolfram. com/GaussianFunction. html. (1999)

59. Neter, J., et al. Applied linear statistical models. (1996)

60. Montgomery, D.C., Peck, E.A. and Vining, G.G. Introduction to linear regression analysis. Wiley. (2012)

61. Kutner, M.H., Nachtsheim, C. and Neter, J. Applied linear regression models. (2004)

62. Wada, Y., et al. Comparison of the methods for profiling glycoprotein glycans—HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study, Glycobiology, 17, 411-422. (2007)

63. Hirabayashi, J. Lectin-based structural glycomics: glycoproteomics and glycan profiling, Glycoconjugate Journal, 21, 35-40. (2004)

64. Josefsson, S. The base16, base32, and base64 data encodings. (2006)

65. North, S.J., et al. Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity, Journal of Biological Chemistry, 285, 5759-5775. (2010)

66. Påhlsson, P., et al. Biochemical characterization of theO-glycans on recombinant glycophorin A expressed in Chinese hamster ovary cells, Glycoconjugate Journal, 11, 43-50. (1994)

67. Chintalacharuvu, K.R., Gurbaxani, B. and Morrison, S.L. Incomplete assembly of IgA2m (2) in Chinese hamster ovary cells, Molecular immunology, 44, 3445-3452. (2007)

68. McCloskey, D.N. and Ziliak, S.T. The standard error of regressions, Journal of Economic Literature, 97-114. (1996)

69. Shen, S., et al. Terminal N-linked galactose is the primary receptor for adeno-associated virus 9, Journal of Biological Chemistry, 286, 13532-13540. (2011)

70. Srikrishna, G., et al. N‐Glycans on the receptor for advanced glycation end products influence amphoterin binding and neurite outgrowth, Journal of neurochemistry, 80, 998-1008. (2002)

71. Nightingale, T.D., et al. A mechanism of sialylation functionally silences the hyaluronan receptor LYVE-1 in lymphatic endothelium, Journal of Biological Chemistry, 284, 3935-3945. (2009)

72. Valencia, J.C., et al. Sialylated core 1 O-glycans influence the sorting of Pmel17/gp100 and determine its capacity to form fibrils, Journal of Biological Chemistry, 282, 11266-11280. (2007)

73. Shibata, T.K., et al. Identification of mono-and disulfated N-Acetyl-lactosaminyl oligosaccharide structures as epitopes specifically recognized by humanized monoclonal antibody HMOCC-1 raised against ovarian cancer, Journal of Biological Chemistry, 287, 6592-6602. (2012)

74. Ashikov, A.M. Cloning and Functional Characterization of Human Nucleotide-sugar Transporters. (2006)

Table 2.1. Proprietary Raw Data File Formats

Vendor	Bruker	Thermo	Waters	MDS/Sciex	Agilent
Instrument Data Acquisition Software	N/A	*XCalibur*	*MicroMass MassLynx*	*Analyst, AnalystQS*	*MassHunter*
Raw File Type	.baf	.Raw file	.Raw directory	.wiff file	.d directory
Raw-to−mzXML Convertor	*CompassXport*[24]	*ReAdW*[24]	*MassWolf*[25]	*mzWiff*[26]	*Trapper*[27]

Table 2.2. Parameters for Spectral Feature Assignment and Extraction

Parameters	Comments
glycomics MS data file	mzXML file
list of glycan structure candidates	Predefined list of glycan structure in XML file
Coverage	Accumulated coverage of isotopomer prevalences. Range from 0~1
minimal charge state in the analysis	Range from 1~4,but should be smaller than or equal to maximal charge in the analysis
maximal charge state in the analysis	Range from 1~4,but should be greater than or equal to minimal charge in the analysis
adduct identities	Defined in configuration file: choice of Na, K, H
Derivatives	Defined in configuration file: choice of Me (methyl), Mx ([^{13}C]-methyl) M1 ([$^{2}H_1$]-methyl), M3 ([$^{2}H_3$]-methyl), Et (Ethyl), Ac (Acetyl), Pr (Propanoyl)
end structure for the glycan structures	Defined in configuration file: choice of "none", "derivatized", "reducing", "alditol" (a polyhydroxy alcohol which is formed by reducing an aldose)
base64 encoded format	whether the data segments will be stored in base64 encoded format or in the two column (m/z and abundance) ASC-II format
isotopically enriched element in the derivative	Defined in isotope configuration file: choice of ^{2}H, ^{13}C, ^{15}N, ^{18}O
ion count	whether the ion count information will be saved in the output file or not

Table 2.3. Abundance vs Mass

Structure	Mass	Abundance
(Hex)1(HexNAc)1	534.2892	65888.48
(NeuAc)1(Hex)1	650.3365	10338.63
(Hex)1(HexNAc)1(Hex)1	738.389	7034.836
(HexNAc)2(Hex)1	779.4155	1121.435
(NeuAc)1(Hex)1(HexNAc)1	895.4628	467719.9
(Hex)1(Deoxyhexose)1(HexNAc)1(Hex)1	912.4781	0
(NeuGc)1(Hex)1(HexNAc)1	925.4734	14070.38
(Hex)2(HexNAc)2	983.5153	2923.252
(NeuAc)2(HexNAc)1	1052.537	812.6886
(NeuAc)1(Hex)1(Deoxyhexose)1(HexNAc)1	1069.552	338.8482
(NeuAc)1(Hex)1(HexNAc)1(Hex)1	1099.563	8673.214
(NeuAc)1(Hex)1(HexNAc)2	1140.589	1341.95
(NeuGc)1(Hex)1(HexNAc)2	1170.6	1081.159
(Hex)3(HexNAc)2	1187.615	3273.66
(NeuAc)2(Hex)1(HexNAc)1	1256.636	210713.9
(NeuAc)1(NeuGc)1(Hex)1(HexNAc)1	1286.647	14203.43
(NeuAc)2(HexNAc)2	1297.663	191.3057
(NeuGc)2(Hex)1(HexNAc)1	1316.658	581.2443
(NeuAc)1(Hex)2(HexNAc)2	1344.689	0
(Hex)3(HexNAc)2(HexNAc)1	1432.741	508.849
(NeuAc)2(Hex)1(HexNAc)2	1501.763	104.6306
(NeuAc)1(Hex)2(Deoxyhexose)1(HexNAc)2	1518.778	123.199
(NeuAc)1(NeuGc)1(Hex)1(HexNAc)2	1531.773	97.98406
(NeuAc)1(Hex)3(HexNAc)2	1548.789	288.3268
(NeuAc)1(Hex)2(HexNAc)3	1589.815	0
(NeuAc)3(Hex)1(HexNAc)1	1617.81	1304.686
(NeuAc)2(Hex)2(HexNAc)2	1705.863	550.8455
(NeuAc)1(Hex)3(HexNAc)3	1793.915	2756.261

Table 2.4. Reproducibility of rolling-trapping MS data for *O*-glycans from Pro5 cells using different abundance scaling methods.

			(Hex)1 (HexNAc)1	(NeuAc)1 (Hex)1 (HexNAc)1	(NeuAc)2 (Hex)1 (HexNAc)1	(NeuAc)1 (NeuGc)1 (Hex)1 (HexNAc)1
	full scan	**average**	2.21%	64.47%	31.41%	1.91%
		standard deviation	0.31%	1.25%	1.34%	0.14%
	rolling-trapping	**average**	3.39%	60.74%	33.71%	2.17%
		standard deviation	0.52%	4.33%	3.72%	0.29%
forced = false	**convergence = 0.9**	**average**	0.93%	37.23%	58.04%	3.80%
		standard deviation	0.49%	19.78%	19.05%	1.22%
forced = false	**convergence = 0.95**	**average**	1.05%	39.92%	55.41%	3.63%
		standard deviation	0.61%	22.72%	21.93%	1.40%
forced = false	**convergence = 0.99**	**average**	1.32%	42.56%	52.68%	3.45%
		standard deviation	0.70%	23.33%	22.59%	1.44%
forced = true	**convergence = 0.9**	**average**	1.33%	46.30%	49.15%	3.22%
		standard deviation	0.23%	8.46%	8.15%	0.55%
forced = true	**convergence = 0.95**	**average**	1.39%	48.40%	47.12%	3.09%
		standard deviation	0.20%	7.28%	7.00%	0.48%
forced = true	**convergence = 0.99**	**average**	1.66%	52.60%	42.93%	2.81%
		standard deviation	0.18%	7.33%	7.05%	0.47%

True indicates that the linear regression caclulation was forced to pass through the origin.

Convergency level: removal of points until R^2 reaches specific value.

Table 2.5. Comparison of highly abundant (CHO-Pro5) *O*-linked glycan structures for three different replicates

	Replicate #	(Hex)1 (HexNAc)1	(NeuAc)1 (Hex)1 (HexNAc)1	(NeuAc)2 (Hex)1 (HexNAc)1	(NeuAc)1 (NeuGc)1 (Hex)1 (HexNAc)1
Full scan	**1**	2.36%	65.86%	29.98%	1.79%
	2	1.86%	63.44%	32.63%	2.07%
	3	2.41%	64.11%	31.62%	1.86%
Rolling-trapping, no scaling	**1**	3.96%	56.09%	37.45%	2.50%
	2	2.93%	61.46%	33.65%	1.96%
	3	3.27%	64.66%	30.01%	2.05%
Forced = false, Convergence = 0.9	**1**	1.32%	53.87%	42.06%	2.74%
	2	1.08%	42.46%	52.94%	3.53%
	3	0.38%	15.36%	79.13%	5.13%
Forced = false, Convergence = 0.95	**1**	1.51%	57.66%	38.33%	2.50%
	2	1.28%	47.78%	47.76%	3.18%
	3	0.36%	14.31%	80.13%	5.19%
Forced = false, Convergence = 0.99	**1**	1.77%	58.60%	37.21%	2.43%
	2	1.68%	53.29%	42.22%	2.81%
	3	0.51%	15.79%	78.60%	5.09%
Forced = true, Convergence = 0.9	**1**	1.60%	56.06%	39.75%	2.59%
	2	1.22%	41.76%	53.46%	3.56%
	3	1.16%	41.08%	54.24%	3.51%
Forced = true, Convergence = 0.95	**1**	1.62%	56.73%	39.10%	2.55%
	2	1.27%	43.26%	52.01%	3.47%
	3	1.28%	45.21%	50.26%	3.26%
Forced = true, Convergence = 0.99	**1**	1.86%	61.00%	34.86%	2.27%
	2	1.60%	49.33%	46.01%	3.07%
	3	1.51%	47.48%	47.91%	3.10%

Table 2.6. Prevalence* of major *O*-glycan structures in replicate samples

		Pro5			Lec2		
structure	**monoisotopic mass**	**average**	**standard deviation**	**relative standard deviation**	**average**	**standard deviation**	**relative standard deviation**
(Hex)1(HexNAc)1	534.28	2.29%	0.09%	0.04	83.00%	0.54%	0.0065
(Hex)1(HexNAc)1 (Hex)1	738.38	0.00%	0.00%	NA	2.48%	0.13%	0.054
(NeuAc)1(Hex)1 (HexNAc)1	895.46	63.02%	1.33%	0.021	13.75%	0.67%	0.049
(NeuGc)1(Hex)1 (HexNAc)1	925.47	1.32%	0.03%	0.026	0.00%	0.00%	NA
(NeuAc)1(Hex)1 (HexNAc)1(Hex)1	1099.56	0.81%	0.02%	0.031	0.00%	0.00%	NA
(Hex)3(HexNAc)2	1187.61	0.00%	0.00%	NA	0.76%	0.01%	0.0091
(NeuAc)2(Hex)1 (HexNAc)1	1256.63	30.70%	1.25%	0.041	0.00%	0.00%	NA
(NeuAc)1(NeuGc)1 (Hex)1(HexNAc)1	1286.65	1.86%	0.14%	0.074	0.00%	0.00%	NA

*Prevalence data corresponds to the abundance of each glycan, normalized to 100%

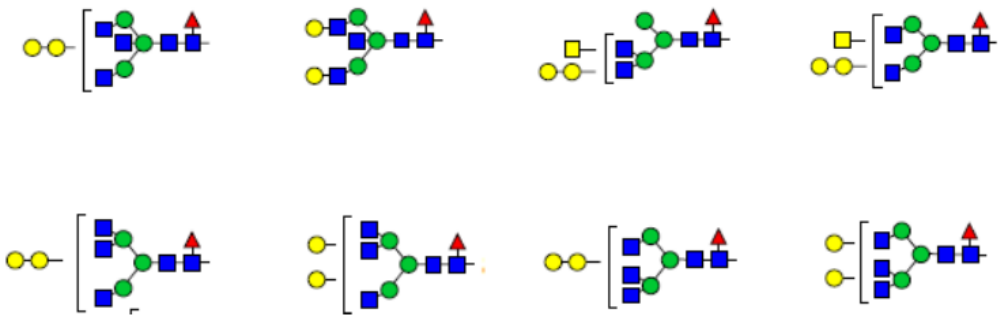


Figure 2.1. An Example of an Isobaric Set: An isobaric set is a collection of isomeric molecules that have a common core structure, presumably due to a common biosynthetic steps. Isobaric sets often occur in glycomics analysis, where a specific type of glycan is being analyzed (e.g., *N*-glycans). Members of an isobolic set cannot be distinguished by 1D mass spectrometry; this requires tandem MS or multiple MS.

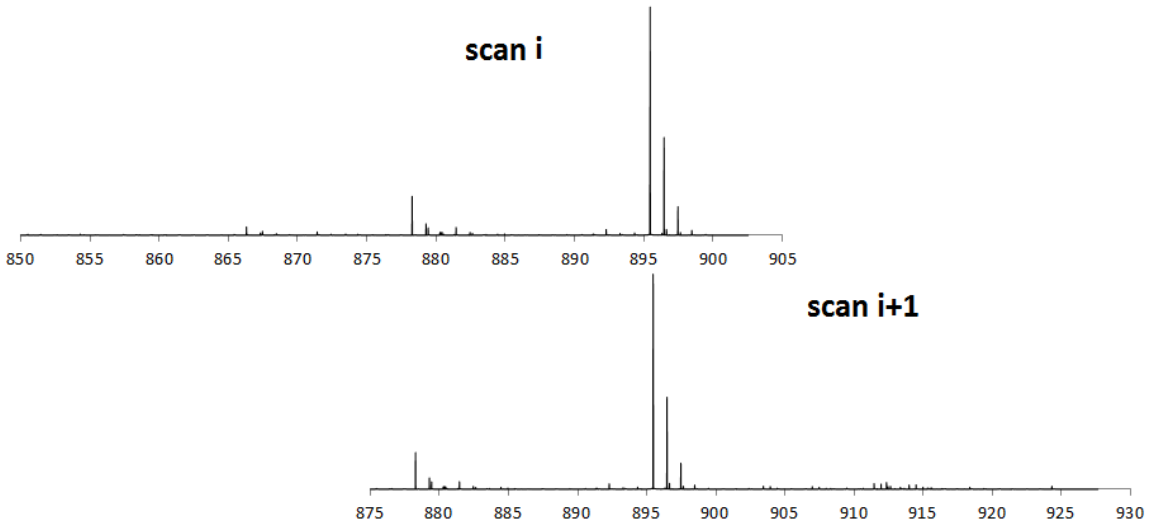


Figure 2.2. Rolling trapping MS. For this method, spectral data are collected as relatively narrow, overlapping segments. This improves the ion statistics (relative to data obtained using full scan methods) by increasing the number of ions counted for each *m/z* value. In this example, a collection of spectra with a width of 50 *m/z* are recorded, incrementing the low mass boundary by 25 *m/z* for each successive scan.

mass	prevalence
534.289	74.65%
535.292	19.27%
536.296	2.38%
536.293	1.69%
537.297	0.44%
535.295	0.39%
535.293	0.31%
535.286	0.28%
537.299	0.19%
536.299	0.10%
536.297	0.08%
536.289	0.07%
538.3	0.05%

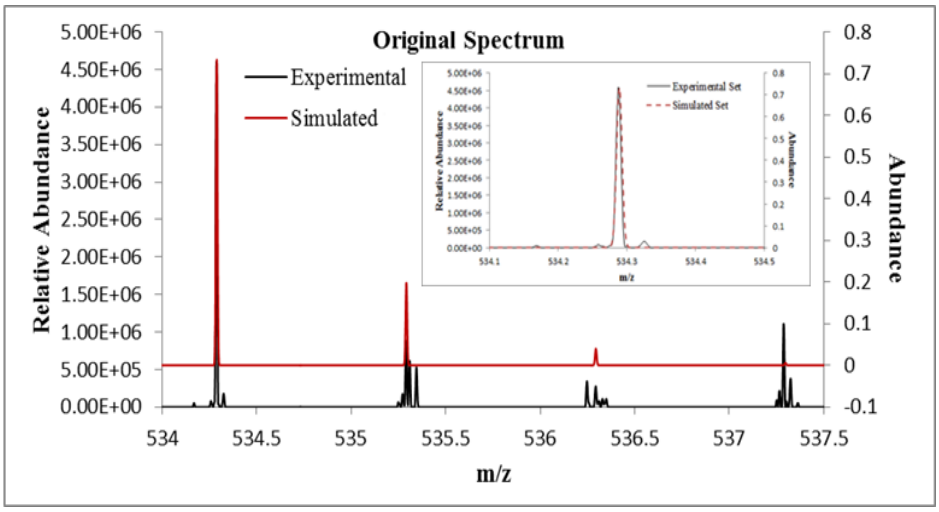


Figure 2.3. Comparison of Experimental vs Simulated Spectrum: The experimental data segment containing the spectral features corresponding to the singly charged, sodiated per-*O*-methylated, alditol form of ($Hex_1HexNAc_1$) was extracted from the mzXML file. The *m/z* range of the extracted data segment is based on calculation of the distribution of isotopologues accounting for 99% of the ions in the isobaric set (listed on the left). The simulated data consists of a sum of Gaussian distributions to generate the vector **S** with components s_i.

$$s_i = \frac{1}{\sigma\sqrt{2\pi}} \sum_j a_j \exp\left(\frac{-\left(m_j - x_i\right)^2}{2\sigma^2} \right)$$

where, a_j and m_j are the abundance and mass of isotopologue j. The simplex algorithm was applied to optimize the peak width and *m/z* offset (δ) for the monoisotopic peak. The optimized values were then used to simulate the signals corresponding to the isotopologues of the glycan ion across entire data segment. Agreement of the simulated and experimental data is illustrated in the inset.

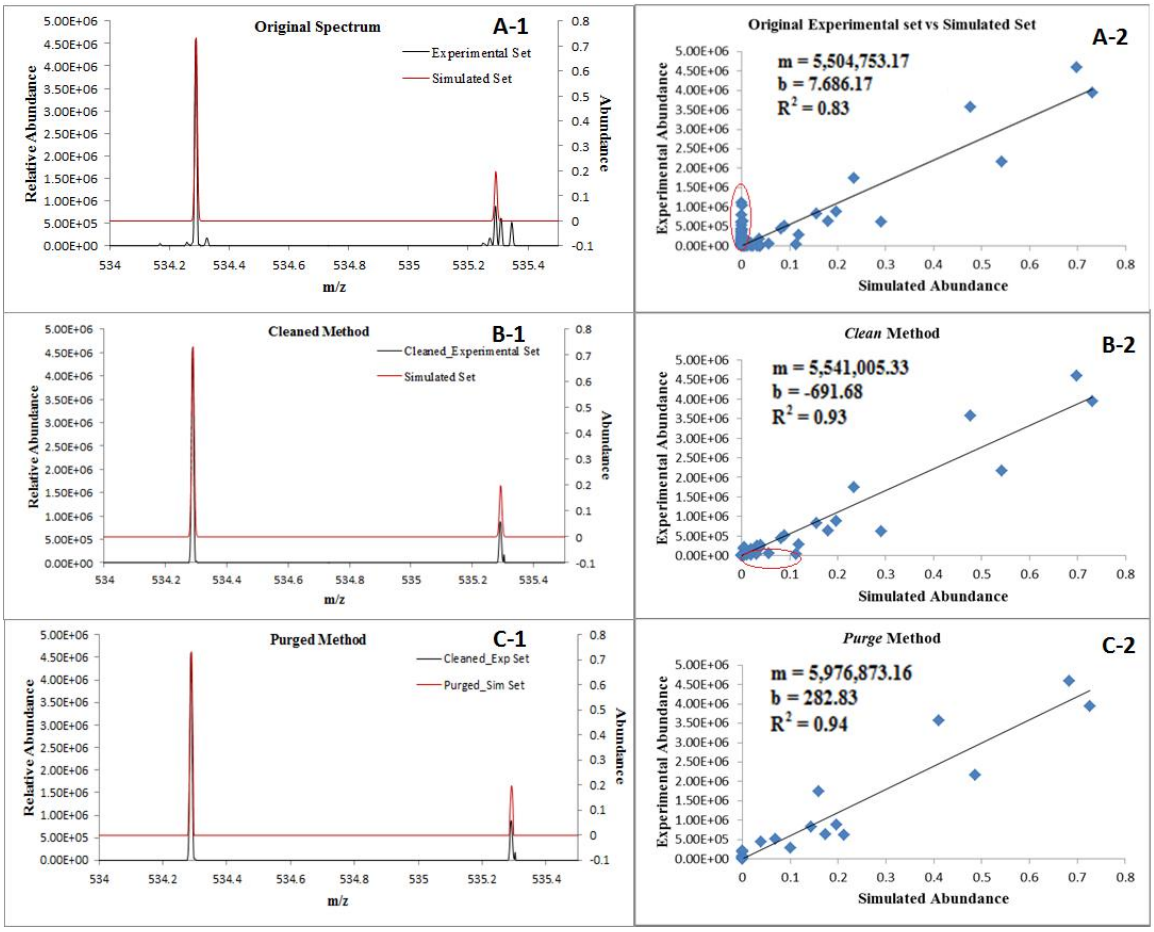


Figure 2.4. Annotation and quantitation of ions corresponding to the singly charged, sodiated per-*O*-methylated, alditol form of ($Hex_1HexNAc_1$)**:** The data segment was extracted as illustrated in Figure 2.3 and the linear regression approach was applied for quantitative analysis. The raw data (panels A-1 and A-2) was modified by sequentially applying the *clean* (panels B-1 and B-2) and *purge* (panels C-1 and C-2) methods to remove noise and contaminant signals. A scatter plot and a linear regression analysis were carried out for each approach, showing that the *clean* method removes signal (represented as data points close to the vertical axis in panel A-2) that are present in the experimental data but not in the simulated data. Subsequent application of the purge method removes signal (represented as data points close to the horizontal axis in

panel B-2) that are present in the experimental data but not in the simulated data. Application of the purge method improves the calculated value of ρ, but often introduces error to the quantitation that is not reflected in this statistic.

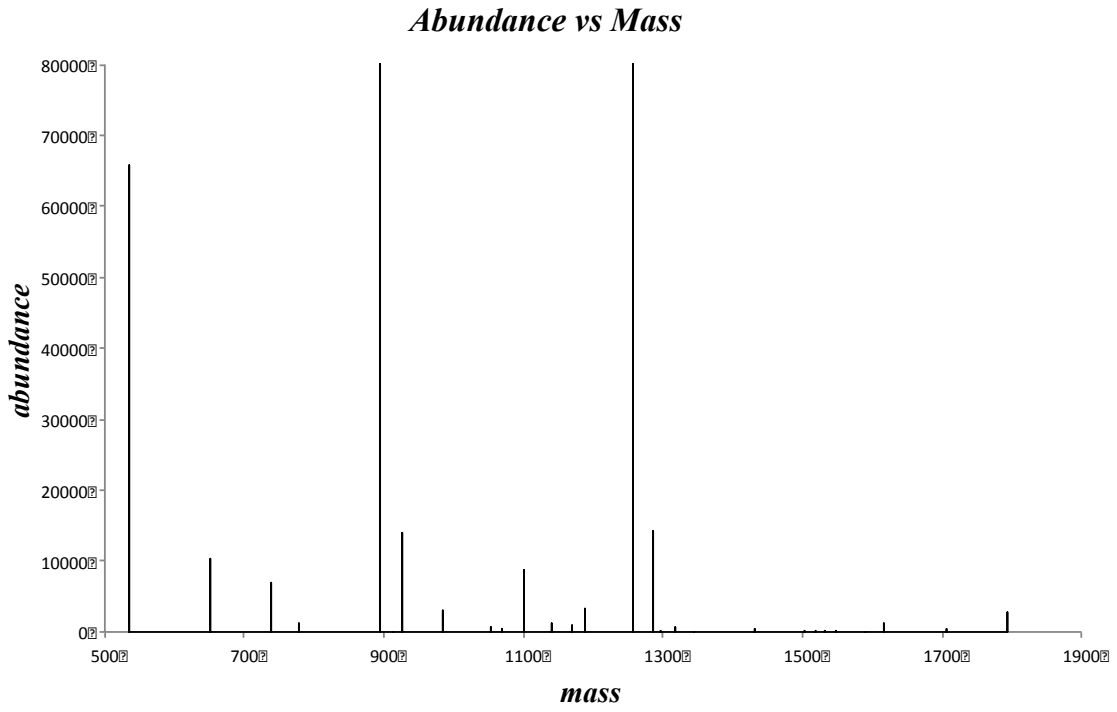


Figure 2.5. The pseudospectrum showing total abundance of each isobaric set of glycans as a function of molecular mass. The threshold for ρ was set to 0, so this image includes data for all glycans in the list, even those whose abundances are so low that they cannot be accurately estimated.

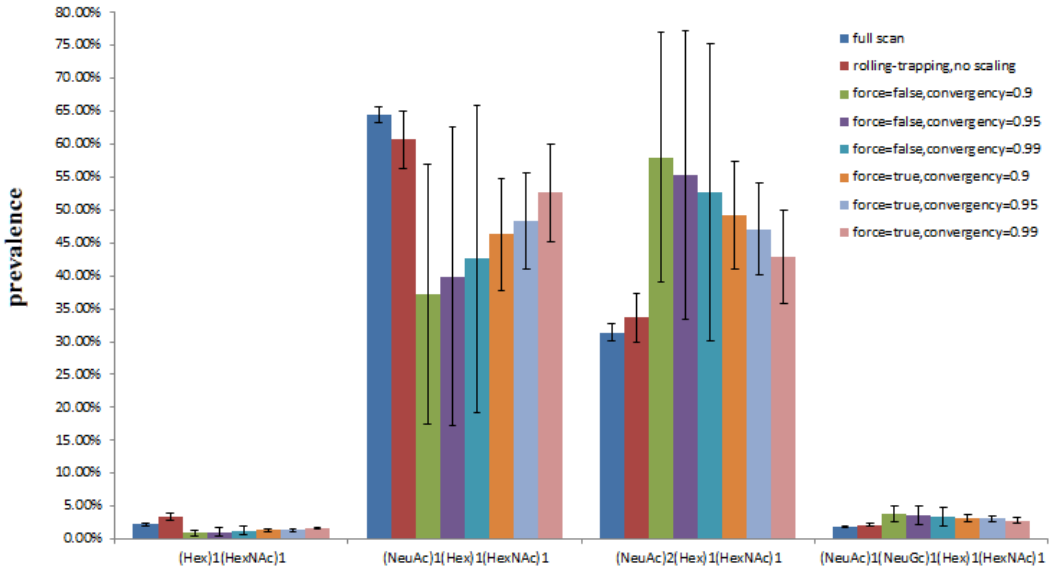


Figure 2.6. Effect of normalizing by comparison of overlapping regions of rolling-trapping scans. Full scan and rolling-trapping data sets (three replicates for each, CHO-Pro5 cell line) were processed through all the modules with different processing parameters to identify the effects of using overlapping segments for abundance normalization for the scans. Data for the four most abundant *O*-glycans in the sample are shown. The best reproducibility (lowest standard deviation, indicated by error bars) was obtained with full scan data (blue bars), which was not normalized. However, this data collection method provided data with poor ion statistics for low abundance ions, compared to rolling-trapping data. Abundance scaling of rolling trapping data performed by comparing overlapping spectral regions increased the replicate-to-replicate deviation. The best results for rolling-trapping data (red bars) were obtained without abundance scaling (other than the default scaling performed during data acquistion by the instrument itself.) Scaling using an algorithm that performs linear regression performed poorly even when outlier points were rejected. As more points were rejected, the results improved, but were never as reproducible as the completely unscaled data.

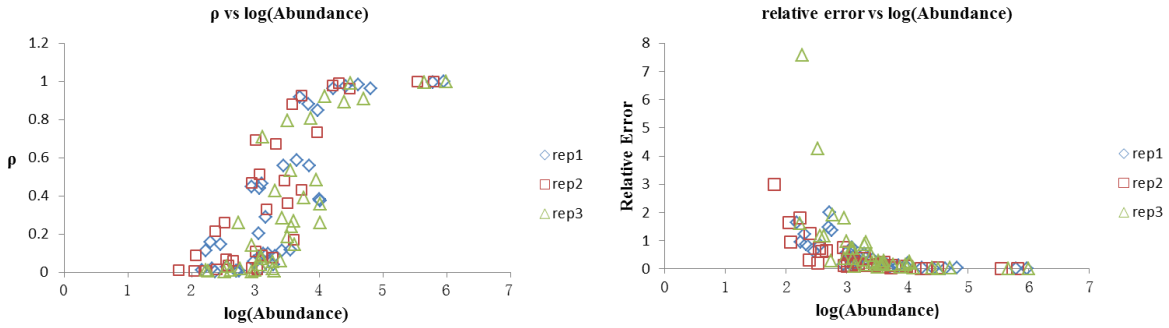


Figure 2.7. Relationships of rho (ρ), ion abundance and relative error. Three replicates of samples from CHO-Pro5 cells were analyzed using the rolling-trapping approach. The raw file was processed through the framework and the cleaned approach was applied for *O*-glycan structure quantification. Statistical parameters, rho (ρ) and relative error obtained during data processing with linear regression are plotted as a function of the log of the calculated ion abundance for each glycan in the candidate structure list. The left panel shows that ρ decreases to unacceptable levels for structures with ion abundances below 10^4 and that the relative error increased for structures with low ion abundances. Thus, quantification of low abundance glycans is not accurate.

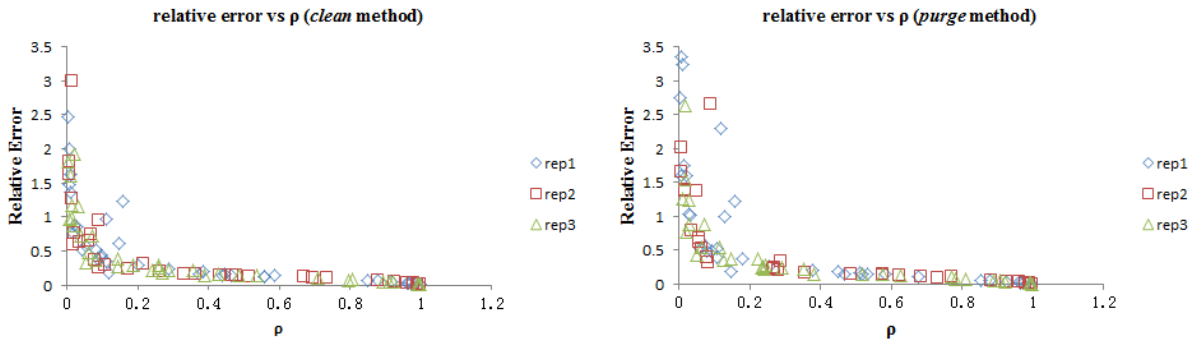


Figure 2.8. Reproducibility of data obtained using the *clean* and *purge* approaches (relative standard deviation for three experiments between experiments vs rho). Three replicates of *O*-glycans from CHO-Pro5 cells were processed with rolling-trapping approach. The raw file was processed using the framework and the *clean* and *purge* data filtering approaches were applied for *O*-glycan structure quantification. As expected, the relative error increases as ρ decreases.

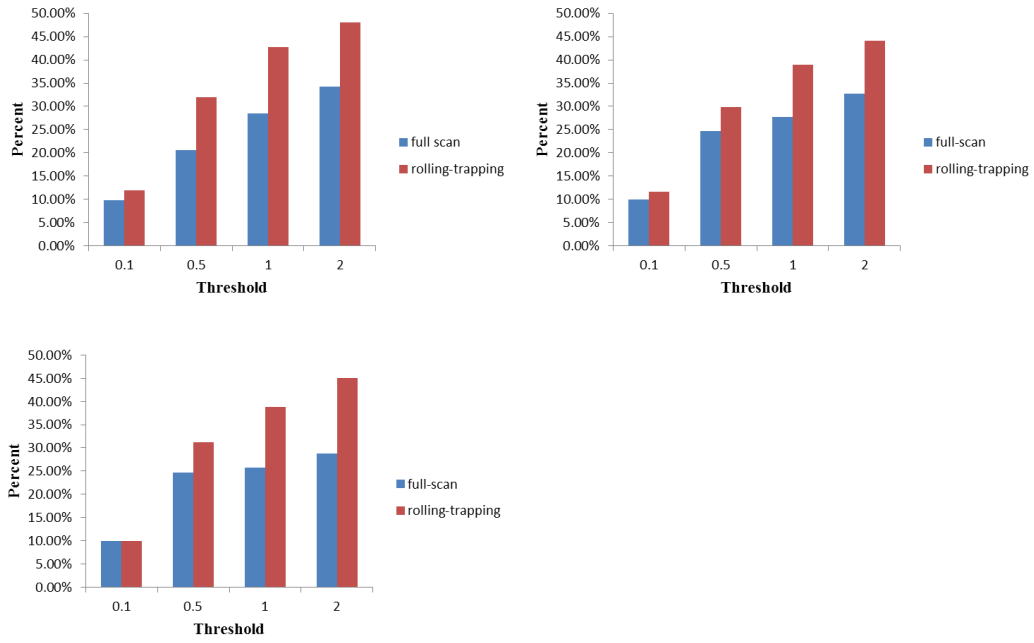


Figure 2.9. Percent of *O*-glycans in the candidate structure list that give rise to "acceptable" data based on different threshold values of the relative standard error. Data were rejected as unacceptable if their relative standard error was above the indicated threshold value (0.1, 0.5, 1.0, 2.0). A relative standard error of 1.0 corresponds to 100% error. These results show that the rolling-trapping method (red bars) provides data with better ion-statistics than does the full-scan method (blue bars).

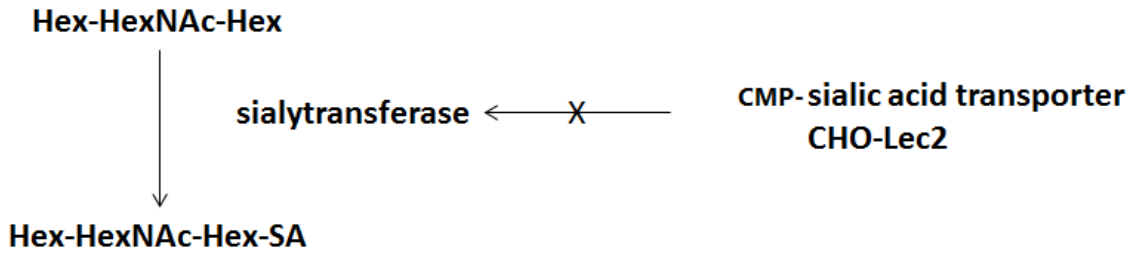


Figure 2.10. CHO-Lec2 cells are deficient in CMP-sialic acid transporter activity.

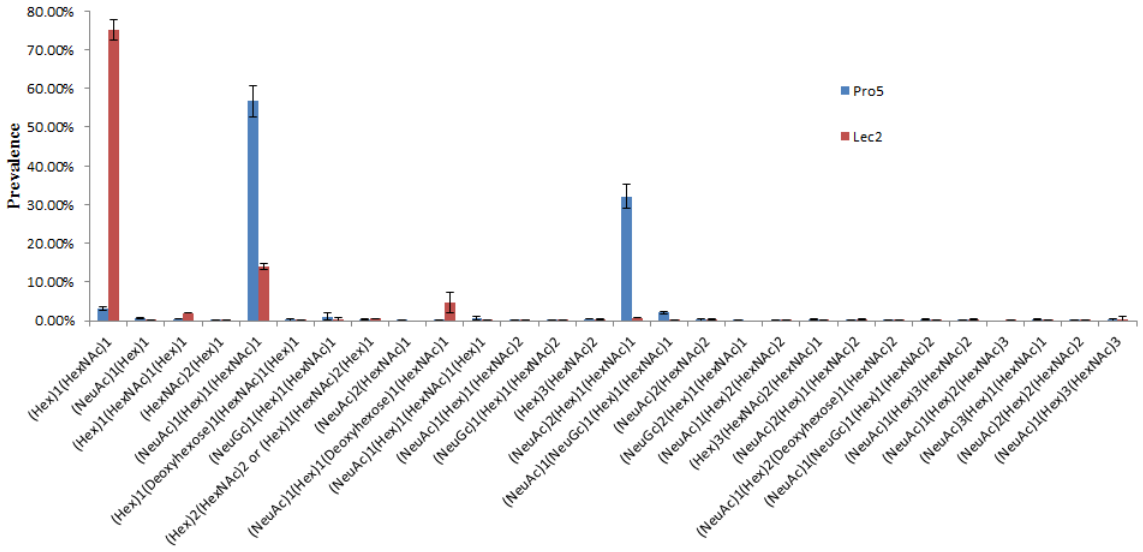


Figure 2.11.A. Prevalence Analysis of Pro5 O-glycans (Overview)

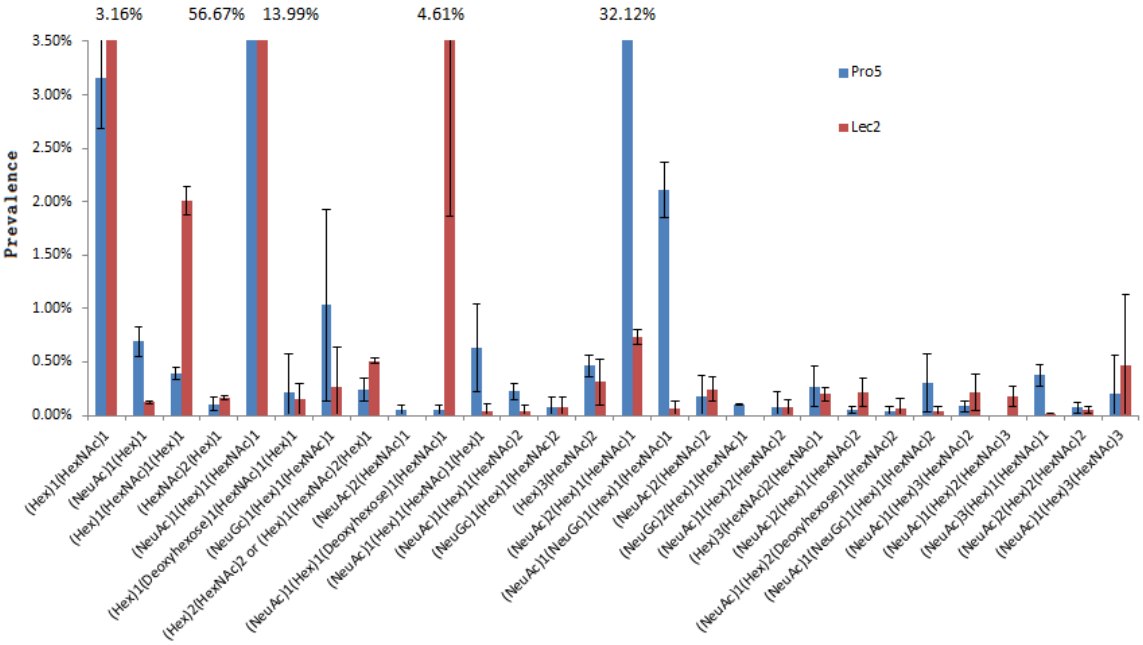


Figure 2.11.B. Prevalence Analysis of Pro5 *O*-glycans (Zoom-In)

Figure2.11. Glycan profiling of *O*-glycans in Pro5 and Lec2 cells. Three replicate analyses of samples from the CHO-Pro5 and Lec2 cell lines were performed using the rolling-trapping approach. The raw files were processed using the modular framework and the *clean* method was

applied for *O*-glycan structure quantification as well as linear regression approach. Prevalence values were calculated based on all *O*-glycan structures in the candidate structure list. Standard deviations were also calculated to show the difference among three replicates. The top and bottom panels show the same data, except the y-axis (abundance) is magnified in the bottom panel. As expected, low-abundance glycans give data with poor statistics, and so their quantification is much less trustworthy than that of the high abundance glycans. In general, the abundance of sialylated glycans is lower in the Lec2 cells, as expected.

CHAPTER 3

ADDITIONAL EVALUATION OF FRAMEWORK IN PERFORMANCE AND GLYCOMICS DATA ANALYSIS

In previous sections, the performance evaluation of the platform was evaluated by processing an MS data file generated using *O*-glycans from CHO-Pro5 and Lec2 cells. Additional performance evaluations were carried out using data sets generated from other biological sources, including human derivative embryonic (denoted as hDE) and human embryonic stem (denoted as hES)[1,2] cells, mouse brain cells and wild type and GalT-knock out (denoted as MEF vs MEF-GalT-KO) mouse embryonic fibroblast cells[3,4]. In this chapter, the performance of platform for identification and quantification of *N*-linked glycan was evaluated using hDE/hES and MEF/MEF-GalT-KO as test cases; the CHO-Pro5/Lec2 was used to evaluate the *purge* method for evaluating ion abundances in spectra with low signal-to-noise ratios; other quality control features were evaluated using MS data sets generated from *O*-linked glycan structures from mouse brain cells.

FRAMEWORK PERFORMANCE EVALUATION USING *N*-LINKED GLYCAN DATA ANALYSIS

One of the motivations for this project is to increase the efficiency of spectral feature annotation, glycan structure identification and quantification. Currently this process is usually performed manually to compare the glycan content of different cell lines or different species, which is a painstaking and slow process. Therefore, to evaluate the performance of our modules and platform, two sets of glycomics MS data were processed as test cases: data obtained by MS

analysis of *N*-glycans from hDE/hES cell lines[5,6] and MEF/MEF-GalT-KO cell lines were processed using the platform and then the results were compared to those obtained manually by experienced glycoanalysts.

The samples used for these test cases contain diverse glycan structures, some of which have higher abundances than others. A candidate glycan structure list for each of these samples was created ("high-prevalence glycan structures") and used as input to test the accuracy of our data processing modules (Table 3.1). Another class ("low-prevalence glycans") was also created to test the sensitivity of our modules. Also, these lists were combined and the prevalence values for all the structure candidates were calculated and compared to results obtained manually.

Profile MS data files generated by analysis of *N*-glycans from hDE, hES, MEF and MEF-GalT-KO cells were processed by the platform to carry out the *N*-glycan structure identification and quantification analysis. Based on the glycan structure list provided by the glycomics researchers from the laboratory, seven abundant *N*-linked glycan structures were selected. MS data files in the .raw file format were converted to mzXML format. Then the list of abundant *N*-linked glycan structures was used to process and annotate the MS data. Annotated MS data segments were selected for several charge states of each structure and then analyzed using the semi-automated framework. The results were compared to results obtained manually (Table.3.1, Figure 3.1).

The results show that our modular framework worked very well, effectively identifying and annotating high-abundance glycan structures and providing quantitative information that is

comparable to that obtained by the human analyst (Figure 3.1, Figure 3.2). Two approaches for quantitative analysis, linear regression and ion counting, were evaluated, as introduced in supplemental section. In this case, the linear regression approach was used to do the structure quantification. Although manual data processing can provide results with good accuracy, identification and quantification of glycans may take several hours to days. In contrast, data processing using our analysis framework took less than ten minutes to complete. This significant increase in efficiency will have a very positive impact on the evaluation of high throughput glycomics data.

Similar analysis was carried out for low-abundance *N*-glycans. A short list of these structures was created and used to process the MEF/MEF-GalT-KO MS data sets to test the sensitivity of identification and evaluate the accuracy of quantification by the modular software (Figure 3.3, Figure 3.4). For these low abundance glycan structures, our platform can still identify most of the structures, although the quantification results are quite different than those obtained manually for some structures. Figure 3.4-A shows that a single structure (N5M3N2) quantified as highly abundant by the linear regression and ion counting methods appears to skew all of the other data. Statistics data for this structure shows that at $z = 2$, the value of ρ is about 0.79, which indicates fairly good fitting between the simulated and the experimental spectrum for this structure. However, the values of ρ at $z = 3$ and 4 are far below 0.8, which indicates that the fit is not good at these two charge states. With the application of our framework, the user has the opportunity to evaluate the quality of the data at each specific charge state. Our software

facilitates this process by providing a visualization of the annotated experimental spectrum's fit with the simulated spectrum. A table of statistics from the analysis is also provided so that the value of ρ can be considered as a metric for confidence in the qualitative and quantitative assignments. In the case of the assignment of N5M3N2, our platform shows that this data set should be carefully examined (Figure 3.3). Doing so revealed that the dominant component of the ion cluster at *m/z* 1210 does not correspond to N5M3N2 but rather to another molecule whose mass is 1 Da lower than that of N5M3N2. Evaluation of the graphical representation of the experimental and simulated data revealed a dominant peak at *m/z* 1209.1, which is *m/z* for an ion with a mass of 2418.2 (z = 2). The mass of the disodiated ion of per-*O*-methylated N5M3N2 ($C_{106}H_{187}O_{51}N_7Na_2$) is 2420.2. Thus, the major ions in this cluster correspond to an unidentified molecule in the sample. Methods to automatically identify this type of error are being developed.

The effectiveness of the linear regression approach for quantification was tested by comparing it to ion counting. In this case, the data from selected spectral segments was filtered using the *clean* approach and the total ion counts in the filtered data was determined. Results from MEF/MEF-KO data sets showed that these two approaches provide very similar results, which are also similar to the results based on manual calculation (Figure 3.5).

PERFORMANCE EVALUATION OF ERROR ESTIMATION FOR PURGED APPROACH

In the last Chapter, the performance of the *clean* method was evaluated in the identification and

quantification of *O*-glycan structures from CHO-Pro5 and Lec2 cells. We also developed another approach, the *purge* method, which simulates the way glycomics researchers often evaluate spectral features with very low signal to noise. In these cases, although some of the predicted isotopologue peaks may be missing, the signal from those peaks that can be manually identified are integrated to estimate glycan abundance. The *purge* method mimics this behavior, and the performance this approach was evaluated and compared with that of the *clean* method. Relationships between the relative standard error, Pearson Correlation Coefficient (denoted as ρ) and abundance (amount calculated from linear regression approach) were identified and characterized (Figure 3.6).

The results were similar to those obtained using the *clean* method evaluated in previous Chapter. That is, the relative error decreased when rho increased, indicating that lower relative error corresponds to better fitting between the experimental and the simulated data set for a specific glycan structure at the selected charge state. Highly abundant analytes give rise to signals that can be quantified with low error and achieve higher values of rho, which indicates higher confidence level for identification and quantification. Higher relative errors (caused, e.g. by noise and contaminating molecules) are associated with lower abundance structures, which agrees with the results obtained using the *clean* methods described in the last Chapter.

To further investigate the performance of *purge* method as well to make comparison between the full scan data set and the rolling-trapping data set, the percent of data that are "acceptable" for different threshold values of rho were also calculated for the results using the *purge* method to processing the CHO-Pro5 and Lec2 data sets. In general, the rolling-trapping

data sets have better quality, since the relative standard errors are smaller than full scan data set. Therefore, at the same value of threshold, the rolling-trapping data set provides quantitation estimates in which a higher percentage of the results for individual structures have relative standard errors below the threshold, when compared to results obtained using the full scan data set (Figure 3.7) The superior performance of rolling-trapping data showed good reproducibility among three replicates, which also indicated good reproducibility of O-glycan structure identification and quantification of glycans from CHO-Pro5 and Lec2 cells. The values of rho increased for each specific structure at each selected charge state when the *purge* method was compared to the *clean* method, as removal of peaks from the simulated data sets by the *purge* method provided a better match to the experimental data (Figure 3.8). It should be noted that, although the *purge* method produced data with better statistics, these statistics are often overly optimistic, as the simulated data is "forced" to correlate with the experimental data.

Based on the performance evaluation of purged method in this section and cleaned method introduced and evaluated in previous section, the following conclusions can be summarized:

(1) The *clean* method improves the quality of glycomics data analysis by reducing errors caused by noise peaks or peaks that do not represent the annotated glycan structures.

(2) Compared to the *clean* approach, the *purge* method is dangerous. Although it can point out the possible presence of a glycan, the quantitation associated with an assignment that becomes statistically acceptable only after the *purge* method has been applied is not trustworthy. Therefore, the purge method significantly affects only low-abundance signals with poor signal to noise, providing a means of annotating these spectral features, albeit with

very low confidence.

(3) The results from both approaches can work as tool to evaluate the quality of MS data sets since our analysis confirmed the superior performance of rolling-trapping technique over the full scan method.

(4) Even through purged approach is an approach based on the output of cleaned approach, sometimes the improved effect is not so obvious without looking at the statistics associated with the results.

PERFORMANCE EVALUATION AND DATA QUALITY EVALUATION ON MOUSE BRAIN DATA SET

The quality of full scan and the rolling-trapping MS data sets generated using *O*-linked glycans from mouse brain cells was also evaluated. These data sets were processed using our glycomics framework to carry out the structure annotation, identification and quantification, and the performance of the software and the data quality were evaluated. The *clean* data filtering method and linear regression method for quantification were used for the data processing, as had already been done for the CHO-Pro5 and Lec2 glycans. The results indicated that our platform can identify most of the structures in the list provided by the user. However, the quality of data set was not as good as that obtained with samples from the CHO-Pro5 and Lec2 cells, as indicated by larger standard deviation values for the estimates of glycan abundance (See Figure 3.9). This indicates that the reproducibility of the *O*-glycan structure profiles are not as good and the estimated abundance for each specific structure varies from sample to sample. To further evaluate these results, the abundances of *O*-glycan structures was also calculated for those

structures having standard errors are below the threshold by removal of those structures with high relative standard errors. Here the threshold was set as 1.0 and structures with relative standard errors above that were removed and the profiling study was repeated again. As indicated in Figure 3.10, even for those structures without high relative standard errors, the reproducibility among replicates were still not as good as had been observed for the CHO-Pro5 data sets, which indicates that the poor quality of the mouse brain data set was not as good as CHO-Pro5/Lec2 set.

Glycans giving abundance estimations with smaller relative standard deviations, include both core structures, e.g. (Hex)1(HexNAc)1 and the extended core structure, e.g. (NeuAx)1(Hex)1(HexNAc)1, as well as some more complex structures, e.g. (NeuAc)1(Hex)2(Deoxyhexose)1(HexNAc)2. Glycans giving abundance estimations with higher relative standard errors usually have smaller values of Pearson Correlation Coefficient at some or among all selected charge states. Although the the poor quality of mouse brain *O*-linked glycan data sets was disappointing, it clearly showed the power of our platform for rapidly evaluating the quality of a data set. This process can take a considerable amount of time, when performed manually. Thus, application of our platform to data sets of poor quality can be beneficial and save valuable time for the glycoanalyst.

To further investigate and evaluate the quality of mouse brain data sets, the relationship between relative standard error (RSE), rho and absolute abundances were studied using the same approach that had been applied to the CHO-Pro5 andLec2 data sets, as shown in Figure 3.11. As previously observed, RSE increased as the values in of rho and absolute abundance decreased, consistent with higher data quality for more abundant analytes. The full scan data set obtained

using glycans from mouse brain cells included fewer structures with acceptable rho values, between 0.8 and 1.0, compared to rolling-trapping data sets, which confirms the relatively poor quality of the full scan data set.

To further characters the effects of data quality on standard errors, the fraction of structures giving abundance estimates with relative standard errors between specific threshold values, e.g., 0.1, 0.5, was also studied. This again provided results similar to those obtained with glycans from CHO-Pro5 and Lec2 cells. However, at a very high threshold value for the RSE, the full scan data set exhibited better coverage than the rolling trapping data set. As shown in Figure 3.12, the full scan data set had increased coverage when the threshold was changed from 0.5 to 2. This result is likely due to the poor overall quality of the rolling-trapping data set for mouse brain *O*-glycans, as the rolling-trapping approach usually provides significantly better coverage.

From the above quality control analysis for the *O*-glycan MS data set from mouse brain cells, the following conclusions can be drawn:

(1) The MS data sets for the *O*-linked glycans from mouse brain cells are of poor quality as indicated by the results obtained by processing the data set with our glycomics analysis framework.

(2) Our glycomics data analysis framework provides a useful tool for the quality control and evaluation of MS data sets, rapidly providing key statistical parameters while performing spectral feature annotation and glycan quantification.

REFERENCE

1. Thomson, J. A. et al. Embryonic Stem Cell Lines Derived from Human Blastocysts. Science 282, 1145-1147, doi:10.1126/science.282.5391.1145 (1998).

2. Wobus, A. M. & Boheler, K. R. Embryonic Stem Cells: Prospects for Developmental Biology and Cell Therapy. Physiological Reviews 85, 635-678, doi:10.1152/physrev.00054.2003 (2005).

3. Nasonkin, I.O. and Koliatsos, V.E. (2006) Nonhuman sialic acid Neu5Gc is very low in human embryonic stem cell-derived neural precursors differentiated with B27/N2 and noggin: Implications for transplantation, Experimental neurology, 201, 525-529.

4. Wang, Y., et al. (2007) Rig-I−/− mice develop colitis associated with downregulation of Gαi2, Cell research, 17, 858-868.

5. D'Amour, K. A. et al. Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat Biotech 23, 1534-1541, doi:http://www.nature.com/nbt/journal/v23/n12/suppinfo/nbt1163_S1.html (2005).

6. Van Hoof, D., Mendelsohn, A. D., Seerke, R., Desai, T. A. & German, M. S. Differentiation of human embryonic stem cells into pancreatic endoderm in patterned size-controlled clusters. Stem Cell Research 6, 276-285, doi:DOI: 10.1016/j.scr.2011.02.004 (2011).

Table 3.1. Comparison of identification (system identification compared to human calculated, hES cells)

Structure	Modular System Identification	Manual Calculation
M5N2-Me	5.17%	6.50%
M6N2-Me	6.19%	7.78%
M7N2-Me	13.86%	17.43%
M8N2-Me	22.69%	28.54%
M9N2-Me	19.36%	24.36%
Glc-M9N2-Me	3.14%	3.96%
SAGal2N2M3N2F-Me	4.88%	6.14%
SA2Gal2N2M3N2F-Me	4.21%	5.29%

Table 3.2. Identification of Low Abundance *N*-glycans

Structure	Manual Calculation	Linear Regression	Ion count
MN2-Me	4.64%	7.56%	7.34%
MN2F-Me	2.10%	4.15%	4.86%
M2N2-Me	3.42%	2.13%	2.33%
SAGalNAcNM3N2-Me	2.56%	4.58%	4.85%
GalNM5N2-Me	3.36%	4.29%	4.22%
GalN2M5N2F-Me	9.76%	6.61%	6.19%
Gal2N2M5N2-Me	9.20%	6.02%	6.87%
GalNM5N2F-Me	0.78%	3.53%	3.00%
GalN2FM3N2F-Me	0.34%	4.00%	4.76%
SAGalN2M3N2-Me	4.36%	4.41%	4.34%
GalNM3M3N2-Me	15.00%	7.01%	7.07%
Gal2N2M3N2F-Me	19.15%	7.29%	7.89%
Gal3N2M3N2-Me	5.11%	3.82%	3.51%
GalN3M3N2F-Me	12.43%	4.21%	4.33%
Gal2N3M2N2-Me	5.99%	3.03%	3.83%
N5M3N2=Me	0.00%	23.62%	21.56%
SAGalN2M3N2F-Me	1.80%	3.75%	3.04%

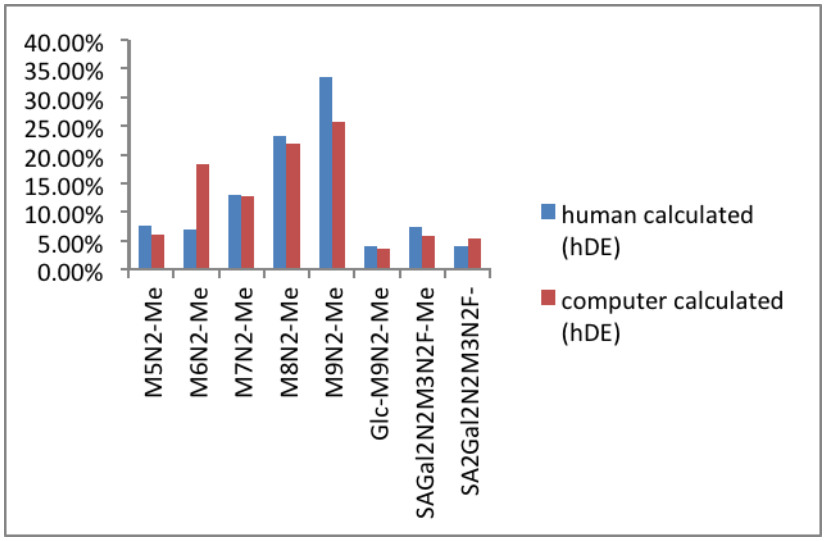


Figure 3.1-A Prevalence Study, hDE cell line

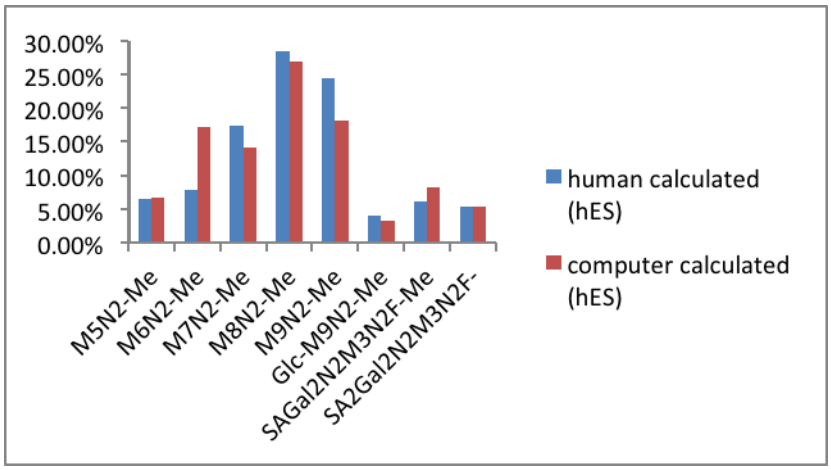


Figure 3.1-B Prevalence Study, hES cell line

Figure 3.1. Prevalenc of N-glycans in hES cells and comparison with the results obtained manually. One glycan sample from hDE/hES cell line was analyzed using the rolling-trapping data acquisition method. The raw file was processed using our modular framework. The *clean* method was applied for signal filtering and *N*-glycans in the candidate structure list were quantified using the linear regression approach.

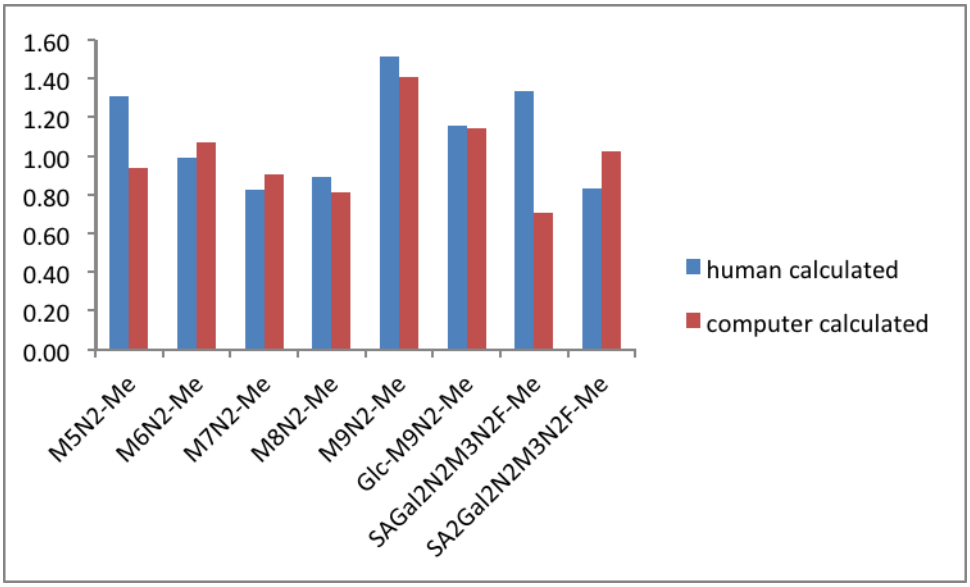
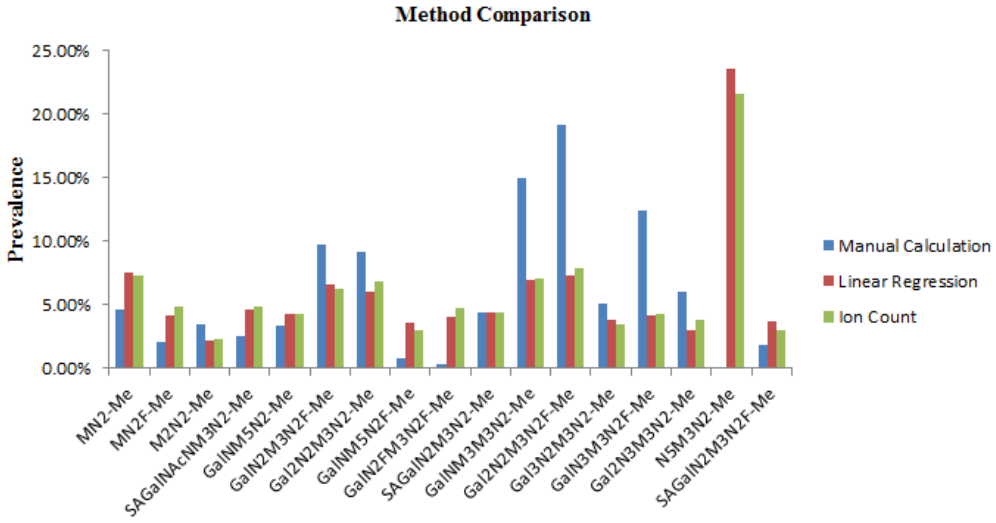
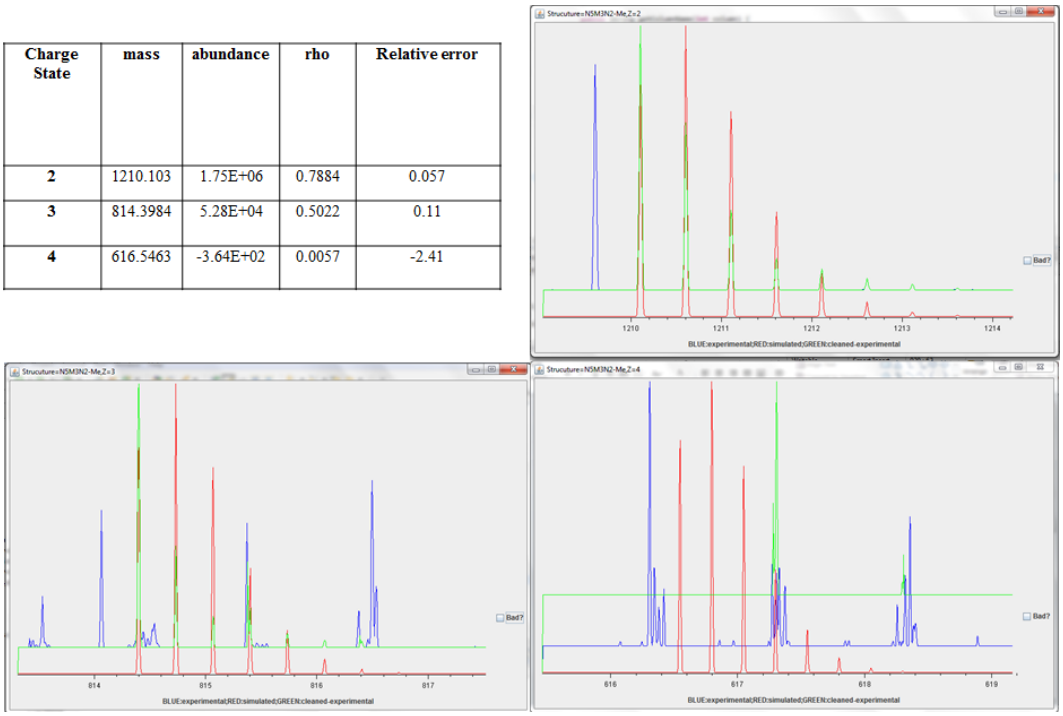


Figure 3.2. Abundance ratios for *N*-glycans from hDE and hES cells and comparison with the results obtained manually. *N*-glycans from hDE and hES cell lines were analyzed using the rolling-trapping data acquisition. The raw file was processed using the modular framework. The *clean* method was applied for signal filtering and *N*-glycans in the candidate structure list were quantified using the linear regression approach. The ratios of glycan prevalence for hDE vs hES cells are shown as determined by the modular data processing framework and manual methods.



A. Profiling study with different approaches

Charge State	mass	abundance	rho	Relative error
2	1210.103	1.75E+06	0.7884	0.057
3	814.3984	5.28E+04	0.5022	0.11
4	616.5463	-3.64E+02	0.0057	-2.41



B. Simulated and Experimental *cleaned* spectrum for N5M3N2.

Figure 3.3. Profiling Study for *N*-linked glycan structures from MEF cells. N-linked glycans from MEF cell line were analyzed with rolling-trapping data acquisition. The raw file was

processed using the modular framework. The *clean* method was applied for signal filtering and *N*-glycans in the candidate structure list were quantified using the linear regression approach and the direct ion counting approach (see Chapter 5). Panel B shows the graphical output of the data visualization module. The spectral segments corresponding to sodiated ions of N5M3N2 with a charge of two, three or four are shown along with the corresponding simulated spectra. The experimental isotopologue patterns for $z = 2$ and $z = 3$ are consistent with a molecule whose mass is one Da greater than that of N5M3N2, indicating that this assignment is incorrect.

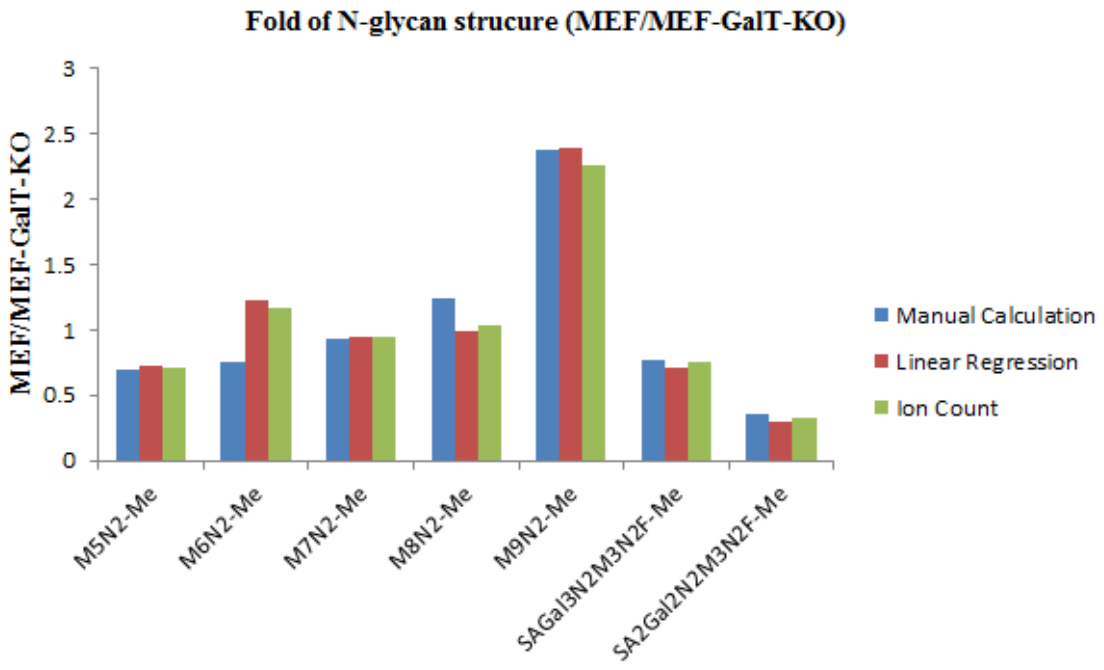


Figure 3.4. Abundance rations for *N*-glycans from MEF/MEF-GalT-KO cells. *N*-glycans from MEF and MEF-GalT-KO cell line were analyzed using the rolling-trapping data acquisition. The raw file was processed using the modular framework. The *clean* method was applied for signal filtering and *N*-glycans in the candidate structure list were quantified using the linear regression approach together with the direct ion counting approach. Only those structures in high abundances are included in this Figure.

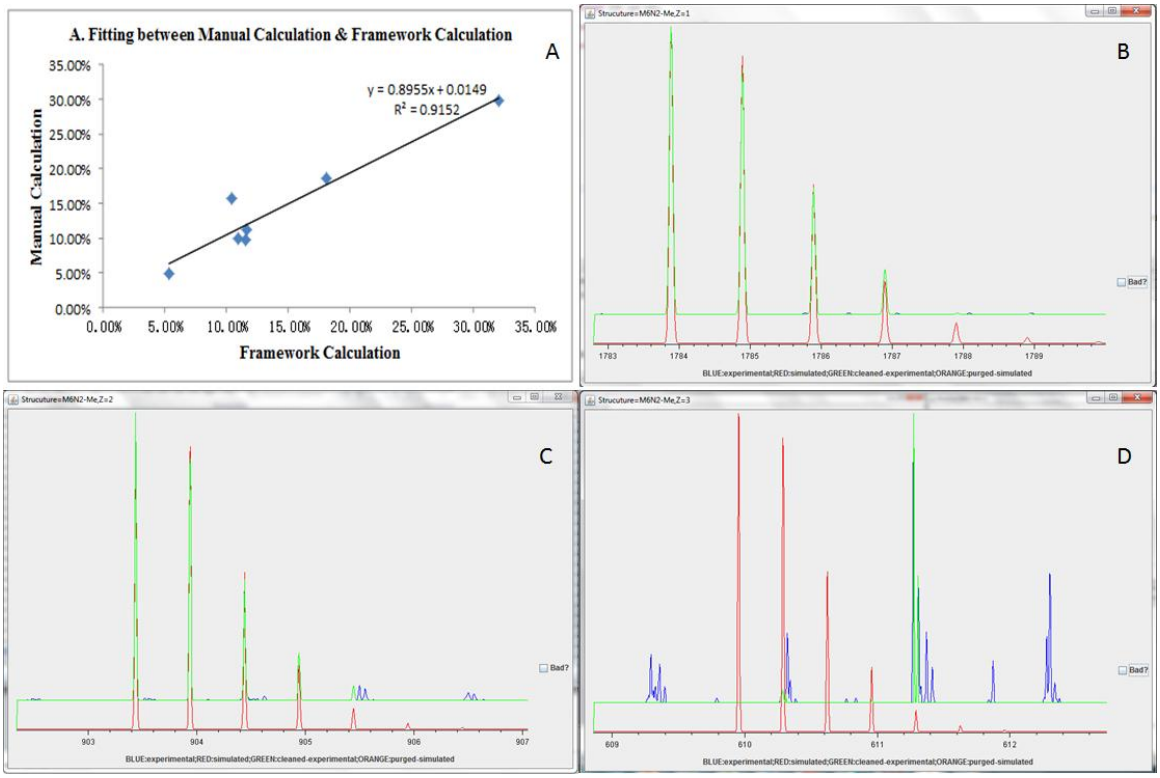


Figure 3.5. Evaluation of abundance estimates for *N*-glycan structures from MEF cells – comparison of results obtained by the modular platform and manual evaluation. MEF *N*-glycans were analyzed using the rolling-trapping data acquisition. The raw file was processed both manually and using the modular framework. The *clean* method was applied for signal filtering and *N*-glycans in the candidate structure list were quantified using the linear regression approach. Results obtained by the modular platform were correlated with results obtained manually (panel A). One point, which represent the structure M6N2-Me (per-*O*-methylated, sodium adduct), is far from the regression line indicating divergence of the results obtained by the two approaches. Panels B, C and D show the experimental data compared to the simulated data for ions having a charge of +1, +2 and +3, respectively. Poor quality statistics for the triply charged ion are confirmed by the graphical data representation in panel D. However, the

triply charged species contributes only a small amount to the ion abundance for this glycan, so this is not the major source of the divergence between the manual and semi automated abundance estimations. The source of this divergence may arise from errors that occurred during manual evaluation, e.g. some noise peaks or peaks that are due to other structures may have been incorrectly assigned to a specific glycan structure at the specified charge state.

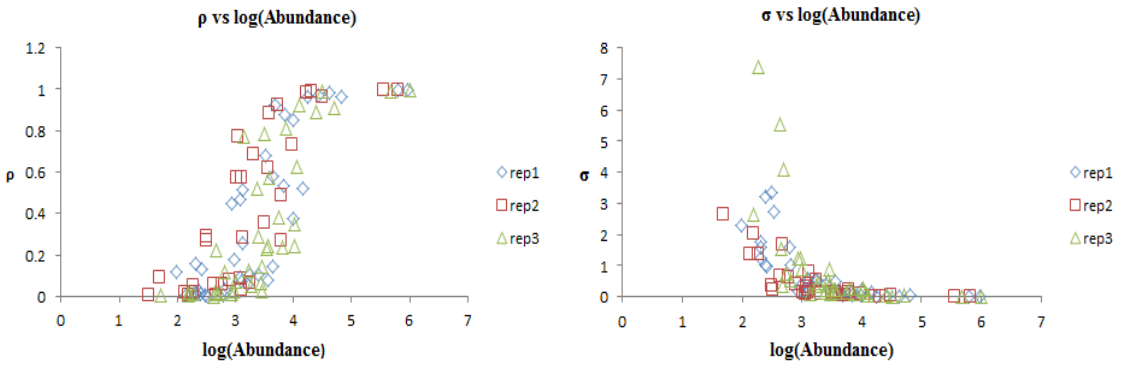


Figure 3.6. Relationship between the relative standard error, Pearson Correlation Coefficient (ρ) and abundance (as calculated using linear regression). *O*-glycans (one replicate) from CHO-Pro5 cells were analyzed using the rolling-trapping data acquisition. The raw file was processed using the modular framework. The *clean* method was applied for signal filtering and *O*-glycans in the candidate structure list were quantified using the linear regression approach. Statistical parameters, rho (ρ) and relative error (σ) were calculated to evaluate their relationship with ion abundance.

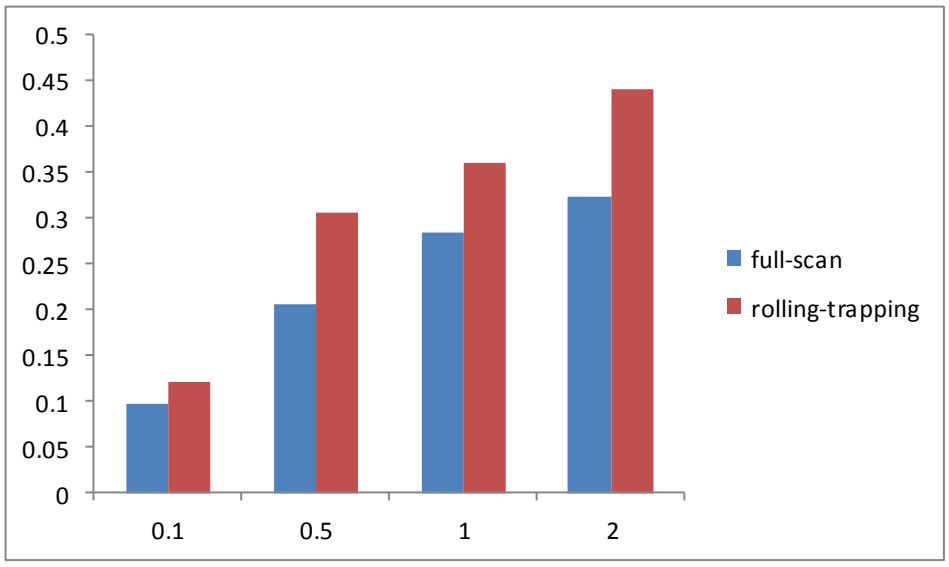


Figure 3.7. Error evaluation of abundance estimates obtained after data filtering using the *purge* method. Percentages of "acceptable" abundance estimates for structure candidates were calculated based on the criterion of relative standard errors below a specified threshold. Different threshold values (0.1, 0.5, 1.0, 2.0) were applied.

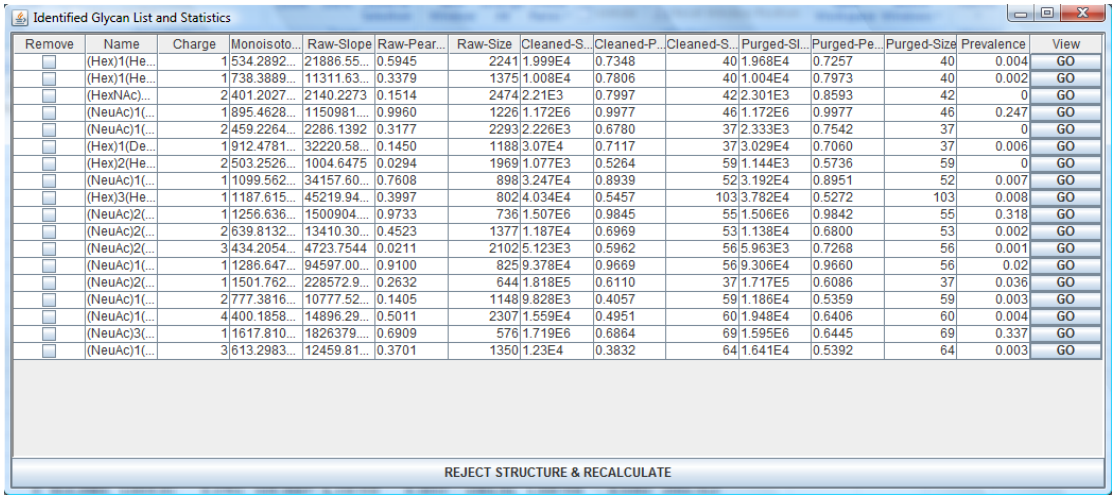


Figure 3.8. Tables of analysis result. Screen shot of the data analysis interface after processing the CHO-Pro5 MS raw data set with *purge* data filtering method and abundance quantification by linear regression. This interface is designed to show the statistics obtained during data processing and allow the user to view the experimental and simulated data for each charge state of each glycan in the candidate structure list.

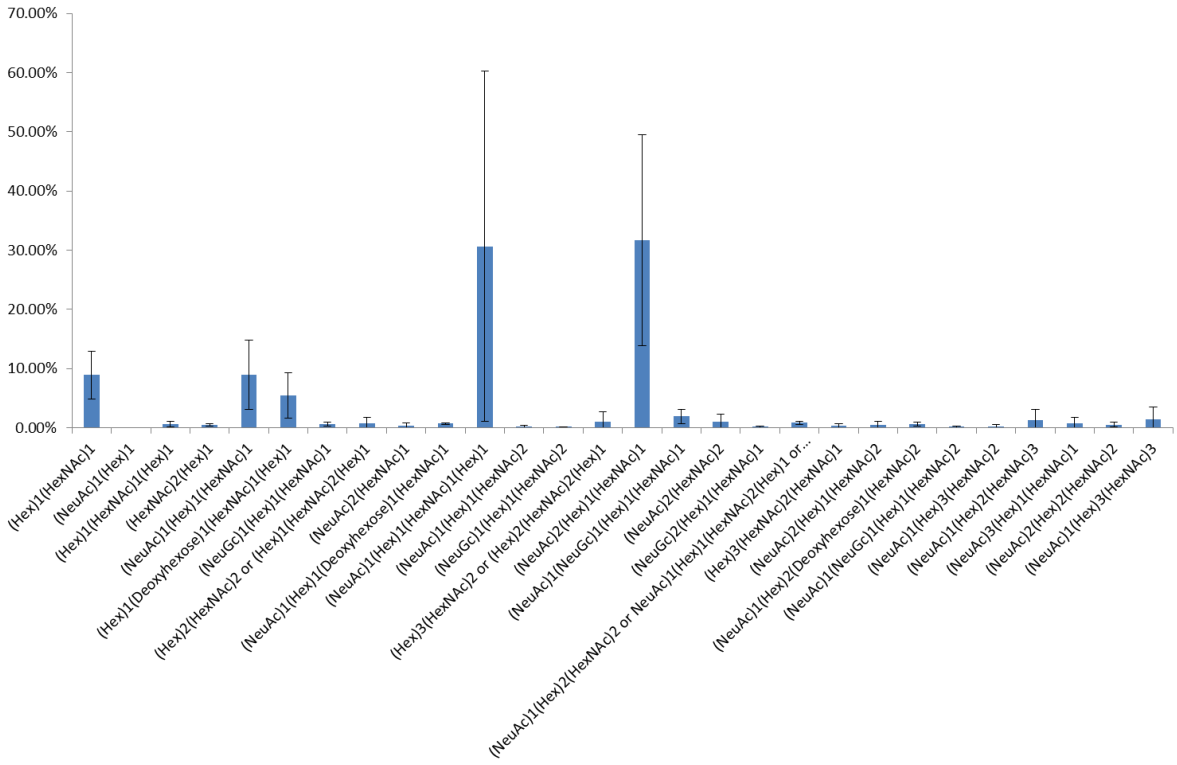


Figure 3.9. Estimated abundances of *O*-linked glycans from mouse brain cells. Three replicate samples were analyzed using the rolling-trapping data acquisition. The raw file was processed using the modular framework. The *clean* method was applied for signal filtering and *O*-glycans in the candidate structure list were quantified using the linear regression approach. Large standard deviations (indicated by error bars) for the three replicates indicate the poor quality of this data set.

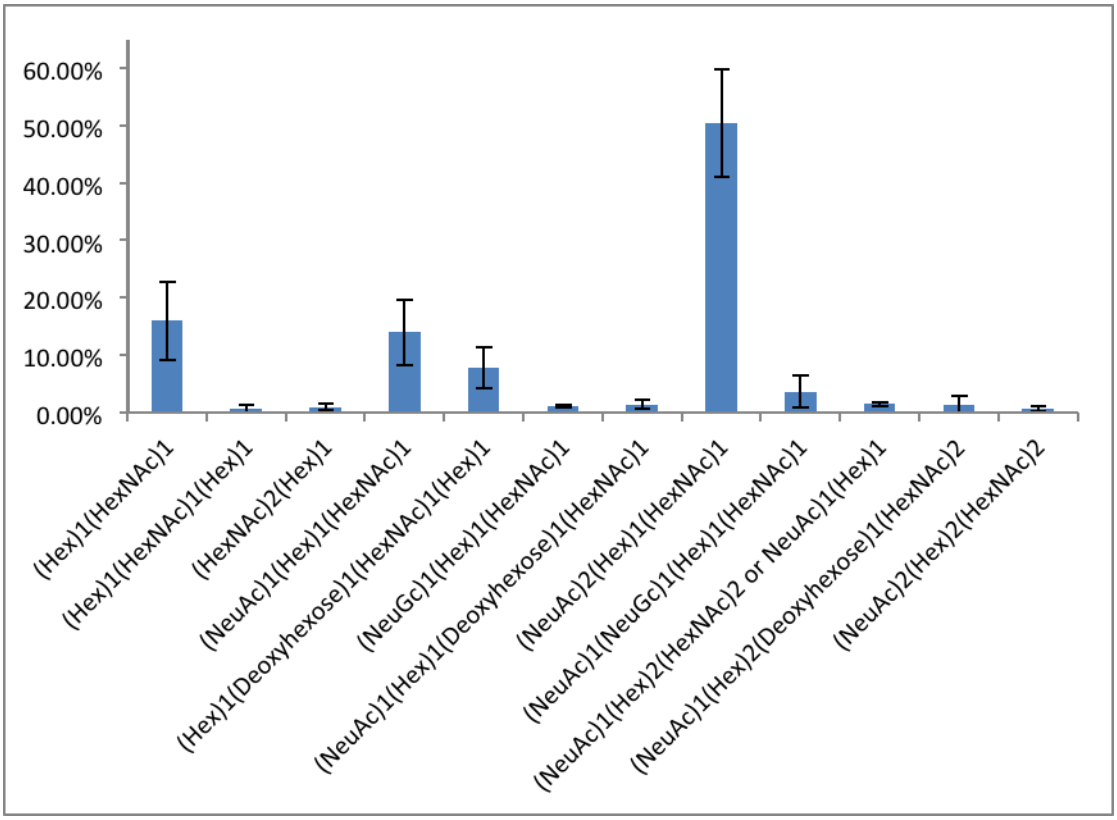
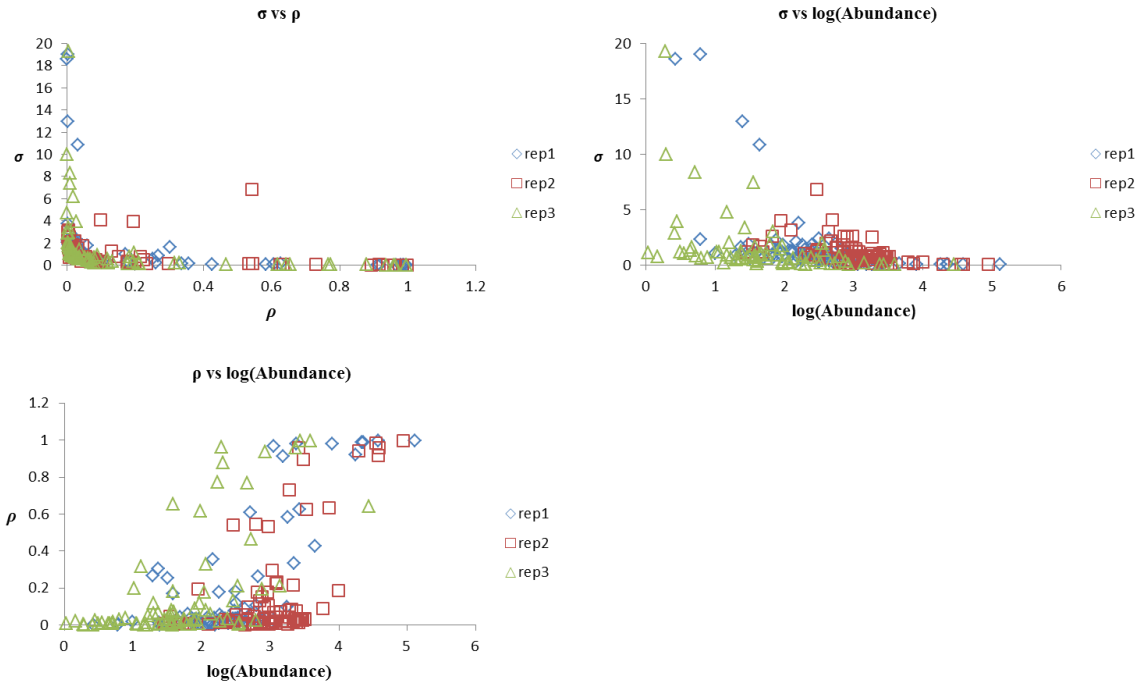
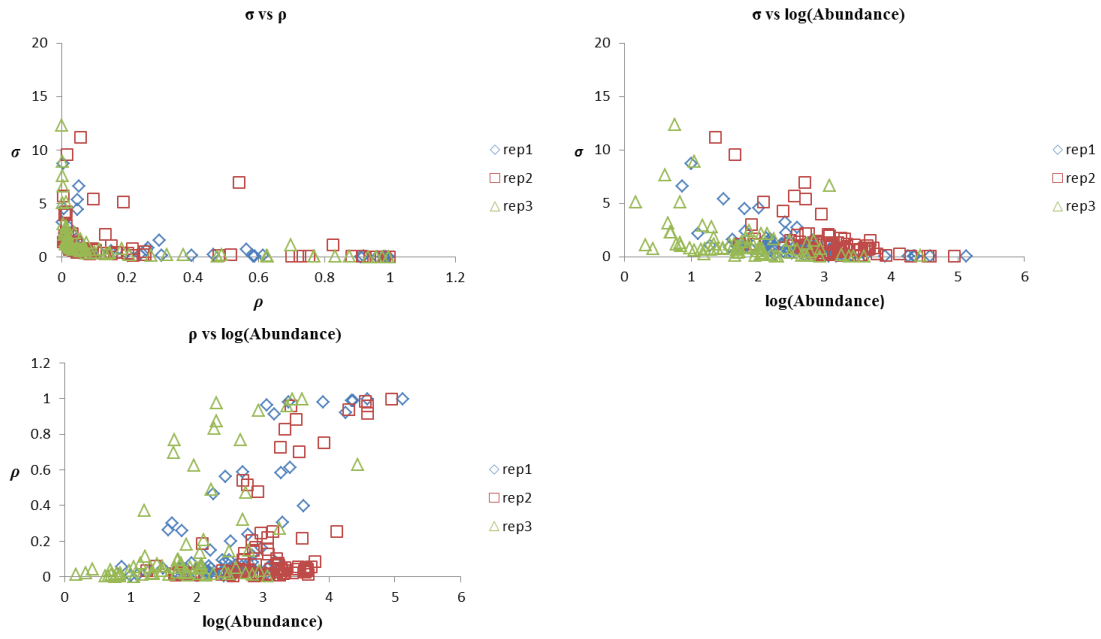


Figure 3.10. Estimated abundances of selected *O*-linked glycans from mouse brain cells. The same data sets shown in Figure 3.9 are illustrated except that data is shown only for *O*-glycan structures having standard errors below a threshold of 1.0. Other structures were removed *before normalization*, showing that inclusion of poor quality abundance estimates has a negative impact on the entire (normalized) data set.



A. Rolling-trapping Data for Glycans from mouse brain - *Clean* Method for Data filtering



B. Rolling-trapping Data for Glycans from mouse brain - *Purge* Method for Data filtering

Figure 3.11. Study of Relative Error, absolute abundance and rho, rolling-trapping, *Clean* and *Purge* methods for mouse brain cells. Three replicate of samples from mouse brain cells

were analyzed using full scan as well as rolling-trapping data acquisition. The two sets of raw files were processed using the modular framework and the *clean* and *purge* data filtering methods were sequentially applied, along with linear regression for *O*-glycan quantification. Statistical parameters, rho (ρ) and relative error (σ) were calculated to study their relationship with ion abundance

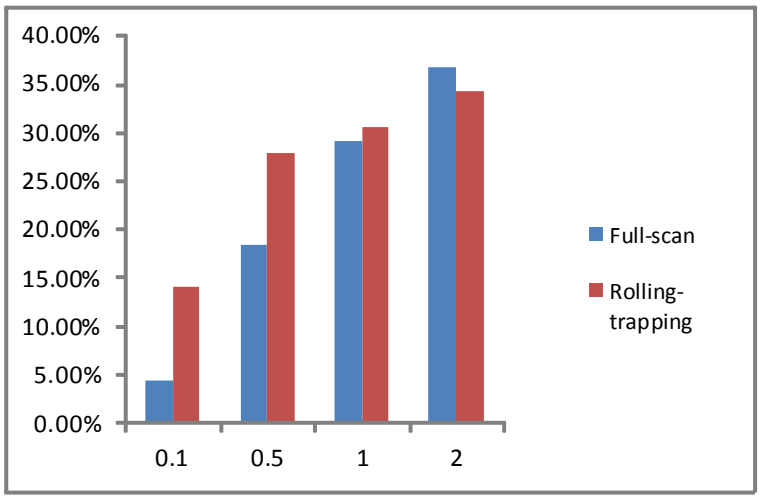


Figure 3.12. Error Evaluation for Estimates of Glycan Abundances in Mouse Brain. The fraction of structure candidates with relative standard errors were below the specified threshold are shown. Threshold values of 0.1, 0.5, 1.0, 2.0 were applied. Data were processed using the *clean* data filtering method together with quantification by linear regression.

CHAPTER 4

CONCLUSIONS

Glycomics is an emerging interdisciplinary research area that is increasingly relevant for biology and chemistry as well as computer science and statistics. Due to the complexity of glycan structure and biosynthesis, sophisticated identification and quantification approaches are often utilized for quantitative glycomics. Rapid progress in experimental methodology (e.g., label-free or isotopic labeling strategies) requires the development of new annotation and quantification software that can keep pace with rapid development of these methodologies, especially for high-throughput glycoanalysis by mass spectrometry.

A semi-automated computational framework to perform quantitative glycomics data analysis was developed to exploit label-free stategies for glycoanalysis. Evaluation of this framework for processing data obtained using rolling-trapping MS methods has shown that it is an effective and efficient tool for annotating and quantifying mass spectral features for predefined lists of glycan structures. The software allows these spectral features to be annotated even in cases where high levels of noise and signals from contaminating molecules are present in the spectra, and provides statistics that indicate the degree to which these unwanted signals decrease data quality. Data processing using this computational framework is semi-automated and requires little human intervention. However, when human intervention is required (e.g., to evaluate data with a high signal to noise ratio), graphical tools are provided to assist the analyst in evaluating the quality of data containing spectral features corresponding to specific glycan structures.

CHAPTER 5

SUPPLEMENTARY MATERIAL

ALGORITHM FOR SCALE NORMALIZATION

Linear regression of the data points from the overlapping sections of a pair of adjacent scans (scan *i* and scan *i+1*) provides a scaling factor f_i. Therefore, from scan 1 to scan *n*, a set of scaling factors { $f_1, f_2, f_3 \ldots\ldots, f_n$ } are calculated. To normalize the abundance signals in scan 2, denoted as A_2, to be consistent with abundance signals in scan 1, denoted as A_1, f_2 is applied to the set of abundances in scan 2, denoted as A_2.

Since $f_2 = \frac{A_2}{A_1}$

the normalized set of abundances in scan 2, denoted as A'_2, is calculated as

$$A'_2 = \frac{A_2}{f_2}$$

However, for scans other than scan 2, the accumulated scaling factors F_i, which is the product of the previous f_i, must be applied.

$$F_i = \prod_{1}^{i} f_i$$

Therefore, $A'_i = \frac{A_i}{F_i}$

ALGORITHM FOR SPECTRAL FEATURE ASSIGNMENT AND EXTRACTION

Glycan structure annotation algorithm:

1) Calculation of isotopologues *m/z* and abundance values to get the m/z range of annotated segments for glycan structures:

i) For each glycan structure candidate in the list, calculate the glycan isotopologue distribution based on the glycan structure, end structure, derivatives, adduct, etc.

ii) The resulting list of isotopologues together with their masses as well as the mass range of the isotopologue distribution [*min*, *max*] based on the parameter "coverage" (i.e., the accumulated coverage from the isotopologue distribution) are used to identify minimal and maximal *m/z* values for the most abundant isotopologues.

iii) Calculate low end and high end of the m/z range from composition data at each charge state

For each charge state *z*, based on the mass range of the abundant isotopologues [*min*, *max*]

$$low\_end = \frac{z * a + \min - d}{z}$$

$$high\_end = \frac{z * a + \max + u}{z}$$

d is denoted for downward extension of peak distribution,

u is denoted for the upward extension of peak distribution

a is the mass of the ionizing species (e.g., Na+).

(a = 39.0983 for K+, a = 22.9897697 for Na+, a = 1.0078250 for H^+)

iv) Calculate the center *c* location of annotated data segment

$$c = \frac{low\_end + high\_end}{2}$$

v) Find the scan where *c* is in or near the center of scan, in which the spectral feature corresponding to the structure is likely to be found

For each scan

a) read the range of m/z distribution [*low_MZ, high_MZ*] as well as the scan number *i* directly from mzXML file

b) calculate center s of each scan

$$s = \frac{low\_MZ + high\_MZ}{2}$$

c) calculate the distance between *s* and *c*, denoted as *delta*

$$delta = |s - c|$$

d) go through all the scans to identify the scan which has the minimum value of delta

2) Pick the scan corresponding to minimal delta and use the Java Random Access Parser (JRAP) application programming interface (API) to get the set of MS data points (*m/z* and *intensity) and* extract a data segment whose *m/z* range is from *low_end* to *high_end* for to annotate this feature annotation as data corresponding to the particular glycan structure.

3) Digitization for annotated structure data set :

For each scan, initialize the optimal gap size as minimal value of step size and find the optimal gap size or step size between data points in the set of { *m/z*, *Abundance* }. The optimized gap size is approximately equal to or slightly smaller than the minimal value of gaps between points within the original set of data points.

The optimized gap size is calculated as follows:

i) Calculate the set of distances between neighbor points, that is, from the set of *m/z* values $\{(m/z)_i\}$, to obtain a set of gaps between each pair of neighbor points among the whole data set, $\{d_1,d_2,d_3,\ldots\ldots,d_n\}$,in which $d_i = (m/z)_i - (m/z)_{i-1}$

ii) \Get the minimal value of *gap_min* from the set of $\{d_1,d_2,d_3,\ldots\ldots,d_n\}$,

where *gap_min* = MIN$\{d_1,d_2,d_3,\ldots\ldots,d_n\}$

iii) Initialize the optimized gap as the value of *gap_min*,

optimal_step=gapmin

iv) Count the number of points, *num*, the optimized gap is applied as follows:

optimal_step=gapmin

num=0;

for each pair of neighbor points:

When $d_i<1.5*optimal\_step$, which corresponds to two points that are close to each other,

num++;

In the other situation, which corresponds to two points separated by a large gap

num+= pointgap/optimal_step,

optimal_step=((m/z)[i]-(m/z)[0])/num

Apply the interpolation algorithm to the set of assigned data segment and fill in the points of {*m/z*, *0*} to ensure a uniform gap size between adjacent points.

4) Apply Java DOM and DOM4j package to parse the information from the input *mzXML* file and create new *XML* file as intermediate file to keep the processed information together with the input parameters

To implement this module to do the structure annotation, one just needs to run the module selection interface and choose the "Structure Annotation" module, and the interface for that module will be accessible. After setting up the values for all the parameters required for the data processing, the interface will appear as shown in the following Figure:

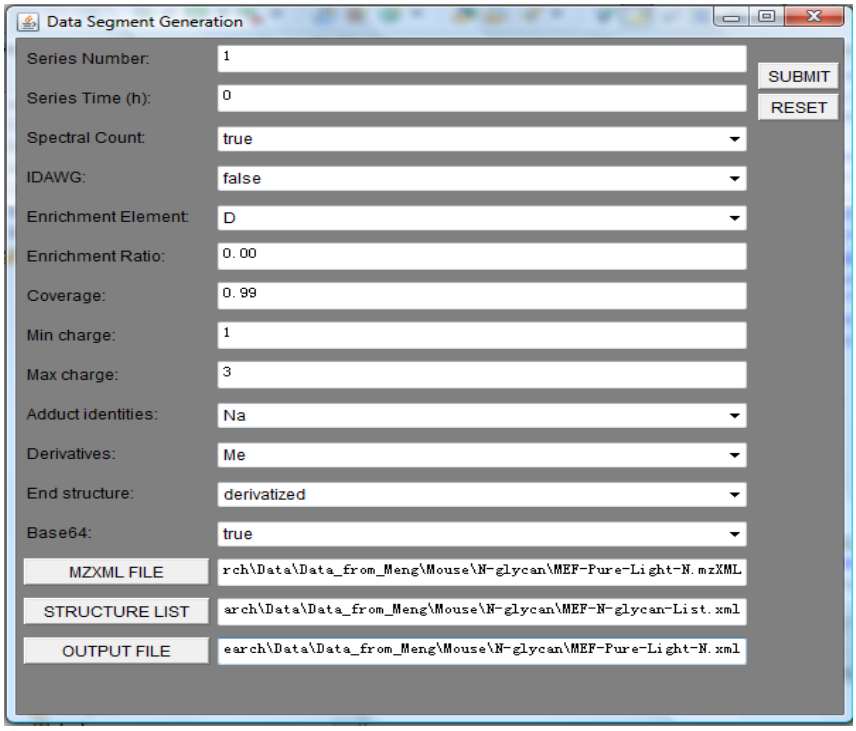


Figure 5.1. Application Interface for Spectral Feature Assignment and Extraction

ALGORITM FOR GLYCAN STRUCTURE CANDIDATE IDENTIFICATION AND QUANTIFICATION

For each glycan structure from the candidate list at each charge state

(1) Calculate the monoisotopic mass to get subset of spectral data points corresponding to the monoisotopic peak from the experimental data set. The algorithms for this step are as following:

Calculate the monoisotopic mass, which is expressed in the unit of daltons (Da). The monoisotopic mass is defined as the sum of the number of atoms of that element times the mass of the most abundant isotope for each element.

(i) Calculate the monoisotopic mass:

$$M = \sum_{1}^{n} n_i * u_i$$

M, monoisotopic mass; n_i, number of atoms of element *i*; u_i, mass of the most abundant isotope of element *i*

(ii) Find the point **P**{*m/z*, abundance} in the annotated data set where the *m/z* is equal or nearest to the monoisotopic mass and get the *m/z* and intensity values at that point, denoted as {*M',A*} here.

(iii) Extend the selection of points to both ends of the *m/z* range in the data set until the abundance for the point is equal or nearest to half of the abundance at point **P**. This

provides a set of data points whose abundance values are within the range of [0.5*A, A], which has a *m/z* range of [m_1, m_2]

(iv) Take the distance between both ends in this range as the initial value for peak width and the difference between monoisotopic mass and the center of the picked peak as the initial value for *delta*. That is

$$pw = m_2 - m_1$$

$$delta = 0$$

(2) Simulate the monoisotopic peak signal.

Typically, in mass spectrometer, a single ionic species is detected as a peak with a Gaussian line shape. Since points within monoisotopic peak in MS data are spread as Gaussian distribution, one can then use this distribution to simulate the monoisotopic peak signal and compare it with the experimental signal.

The monoisotopic signal distribution is calculated as a vector **S**, whose elements s_i are

$$s_i = a * \exp\left(\frac{-(m - x_i)^2}{2\sigma^2} \right)$$

a is the abundance of monoisotopic peak

m_j is the monoisotopic *m/z*

x_i is *m/z* for data point i,

σ is the standard deviation of the Gaussian function

σ is related to the peak width: $\sigma = 0.4247\ pw$ or $pw = 2.3548$

The values of x_i are taken from the interpolated data segment, as described above

(3) Use Nelder-Mead simplex algorithm to optimize the simulation of monoisotpic peak to get the optimized peak width and delta. The detailed steps to implement this algorithm are as follows:

i) Initialize the values of peak width and delta

ii) Initialize the simulation of monoisotopic peak

iii) Initialize the linear regression between the experimental data segment and simulated monoisotopic peaks and get the initialized R^2

iv) Use Nelder-Mead algorithm, which is a simplex method to find the local minimal set of variables, in this case, to determine the values of peak width and delta at which the maximal value of R^2, is attained.

A mathematic optimization problem can usually be defined as follows:

Given Input: a mathematic function or formula F(x), which is called objective function or cost function, with a set of variables X $\{x_i\}$ and their initial values

Sought Output: a set of variable values X_0, an optimal solution from all possible sets of X, which is usually called search space, a collection of candidate solutions or feasible solutions, such that F(x) reaches its minimization state F_0

In the real world, many problems can be modeled generally into this kind of theoretical problem, which is used to find a local minimum for the set of variables from their function. In our implementation of this algorithm, the input and output are instantiated as follows:

Input: monoisotopic peak simulation function, the linear regression of the experimental abundance data set and the simulated abundance data set and the initial values of peak width and delta

Output: the optimized set of {peak width, delta} at which the negative value of R^2 reaches its minimum

In this case, the optimization process for the two variables, peak width and delta, the simplex is a triangle, and this non-linear programming method is instantiated as a pattern search that compares function values at the three vertices of a triangle. After rejecting the vertex where the negative value of R^2 is largest, a new vertex is generated with a new set of {peak width, delta} to replace the rejected vertex, and a new triangle is then formed for a new iteration. The optimization search is continued in this way and the searching process generates a sequence of triangles in which the negative values of R^2 at the vertices decreasing. The size of the triangles is reduced until the optimized set of {peak width, delta} is found where negative value of R^2 arrives at its minimum. In this way, the maximum of R^2 is reached in an effective and computationally compact process.

(4) Use optimized peak width and delta to simulate whole spectrum distribution for each structure.

The whole spectrum is calculated as a vector **S**, whose elements s_i are

$$s_i = \frac{1}{\sigma\sqrt{2\pi}} \sum_{j \le n} a_j \exp\left(\frac{-(m_j - x_i)^2}{2\sigma^2}\right)$$

a_j is the abundance of isotopomer j in a list containing n isotopomers

m_j is the exact m/z for isotopomer j

x_i is m/z for data point i,

σ is the standard deviation of the Gaussian function

σ is related to the peak width: $\sigma = 0.4247\ pw$ or $pw = 2.3548\ \sigma$

n is the number of isotopologues used in the simulation

IMPLEMENTATION OF LINEAR REGRESSION

To implement the linear regression method, the input data sets will be the set of abundances for both experimental data set *E* and the simulated data set *S*. The implementation of linear regression in this project will follow the following steps:

(1) Calculation of slope *m* and intercept *b*:

(i) Calculate intermediate variables *SumS, SumE, SumSS, SumSE, SumEE*

Where

$$SumS = \sum_{i=1}^{n} S_i$$

$$SumE = \sum_{i=1}^{n} \varepsilon_i$$

$$SumSS = \sum_{i=1}^{n} S_i^2$$

$$SumSE = \sum_{i=1}^{n} (s_i * \varepsilon_i)$$

$$SumEE = \sum_{i=1}^{n} \varepsilon_i^2$$

(ii) Calculate *m* and *b* with the following equations:

$$m = \frac{n * SumSE - SumS * SumE}{n * SumSS - (SumS)^2}$$

$$b = \frac{SumE - m * SumS}{n}$$

(2) Calculation of error *E*:

(i) Calculation of *SSE* and *SST*

$$SSE = \sum_{i=1}^{n} [\varepsilon_i - (m * s_i + b)]^2$$

$$SST = \frac{SumEE - SumE * SumE}{n}$$

(ii) Calculation error *E*

$$E = \frac{SSE}{SST}$$

(3) Calculation of R^2:

$$R^2 = 1 - E$$

(4) Calculation of standard error:

For the experimental abundance data set $E\{e_1, e_2, \ldots\ldots, e_n\}$ and $S\{s_1, s_2, \ldots\ldots, s_n\}$, after we have got the linear relationship $E=m*S+b$ between *E* and *S*

i) calculate the average of two data set E and S to get $\bar{s}$ and $\bar{e}$

$$\bar{s} = \frac{\sum_{i=1}^{n} s_i}{n}$$

$$\bar{e} = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}$$

ii) calculate the distance between s_i and $\bar{s}$ to get $\Delta S = \{ \Delta s_i \}$ as well as $\Delta E = \{ \Delta e_i \}$

$$\Delta s_i = s_i - \bar{s}$$

iii) calculate Δ^2 for ΔS and ΔE as well as the set of functions of ΔS and ΔE to get ΔSS, ΔEE and ΔSE:

$\Delta SS = \{(\Delta s_i)^2\}$

$\Delta EE = \{(\Delta e_i)^2\}$

$\Delta SE = \{(\Delta s_i)*(\Delta e_i)\}$

iv) calculate the error square ES between the simulated set S and experimental set E

$$es_i = [e_i - (m * s_i + b)]^2$$

v) get the sum of error square SES

$$SES = \sum_{i=1}^{n} es_i$$

vi) calculate the standard error *STDE* according to the equation

$$STDE = \sqrt{\frac{SES}{(n-2) * \sum_{i=1}^{n} (\Delta s_i)^2}}$$

INTERPOLATION ALGORITHM

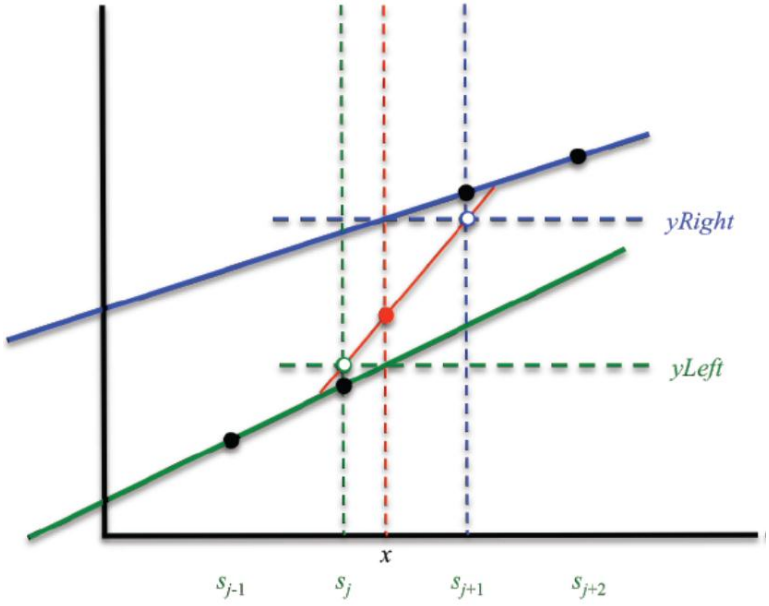


Figure 5.2. Interpolation Algorithm.

This algorithm is designed to calculate the interpolated y - values for an evenly spaced set of *x*-values, given an input set $\{(s_1, t_1), (s_2, t_2), \ldots\ldots, (s_n, t_n)\}$ of Cartesian coordinates with arbitrary spacing between the *s* - values. Consider the small input set of four points in the above Figure. The goal is to calculate the interpolated value of y at an arbitrary *x* (dashed red vertical line). First, an index *j* for the input set is chosen such that $s_j <= x$ and $s_{j+1} > x$. Then, a line (solid green) is drawn through the two points (s_j,t_j) and (s_{j-1},t_{j-1}). The *y*-value on that line at *x* is calculated (at the intersection of the solid green line and the vertical dashed red line.) This corresponds to the value y_{Left} that would be calculated based solely on the two points (s_j,t_j) and (s_{j-1},t_{j-1}). Similarly, a line (solid blue) is drawn through the two points, (s_{j+1}, t_{j+1}) and (s_{j+2}, t_{j+2}). The y - value on that line at x is calculated at the intersection of the solid blue line and the vertical

dashed red line. This corresponds to the value y_{Right} that would be calculated based solely on the two points (s_{j+1}, t_{j+1}) and (s_{j+2}, t_{j+2}). The desired y - value at x is some combination of y_{Left} and y_{Right}. This is calculated by drawing a line (solid red) between the points (s_j, y_{Left}) [open green circle] and (s_{j+1}, y_{Right}) [open blue circle] and choosing the y - value on that line at x to give a function $y = f(x)$. This function has the property that $f(s_i) = t_i$. That is, the curve passes precisely through each point defined in the input set. One advantage of this algorithm is its simplicity. Unfortunately, the first derivative of this function is discontinuous (unlike a cubic spline). This may be acceptable, if all that is required is a reasonable interpolation. Useful data (with more subtle discontinuities of the first derivative) may be generated by making a linear combination of a completely linear, two - point interpolation and the four - point interpolation described here. This requires specification of a weighting parameter $0 < w < 1$, which is implemented by defining the final (red) line as passing through the points

(s_j, $w*y_{Left} + (1 - w)*t_j$) and (s_{j+1}, $w*y_{Right} + (1 - w)*t_{j+1}$).