DEVELOPMENT AND APPLICATION OF AUTOMATED INTERPRETATION TOOLS FOR STRUCTURAL CHARACTERIZATION OF GLYCOSAMINOGLYCANS USING MASS SPECTROMETRY

by

JIANA DUAN

(Under the Direction of I. Jonathan Amster)

ABSTRACT

Glycosaminoglycan (GAGs) are linear chain glycans consisting of repeating uronic sugar and amino sugar copolymers and play major roles in fundamental biological processes. Despite being ubiquitous in cells, the structure of intact GAG chains remains relatively elusive. GAG sequences can be determined using mass spectrometry with ion activation techniques but structure characterization is both time consuming if performed manually and requires a high degree of expertise. The structural analysis step is the bottleneck for higher-throughput methodologies, making it difficult for biological laboratories and clinics to perform routine glycan characterization. The work here is a software solution to the interpretation step, optimizing structures based on highest likelihood while using a genetic algorithm to maximize computation efficiency. This software package is put to the test by 1) determining large but well characterized glycans as well as 2) unknown structures far too complex for manual interpretation.

INDEX WORDS: Mass Spectrometry, Glycosaminoglycan, Software Development

DEVELOPMENT AND APPLICATION OF AUTOMATED INTERPRETATION TOOLS FOR STRUCTURAL CHARACTERIZATION OF GLYCOSAMINOGLYCANS USING MASS SPECTROMETRY

by

JIANA DUAN

PhD, University of Georgia, 2018

BS, Georgia Institute of Technology, 2011

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

2018

DEVELOPMENT AND APPLICATION OF AUTOMATED INTERPRETATION TOOLS FOR STRUCTURAL CHARACTERIZATION OF GLYCOSAMINOGLYCANS USING MASS SPECTROMETRY

by

JIANA DUAN

Major Professor:	I. Jonathan Amster
Committee:	Jeffrey Urbauer
	Ron Orlando

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

DEDICATION

For Ben, Yushu and Yixiang. Family is everything.

For me. You worked hard.

ACKNOWLEDGEMENTS

I'll start by acknowledging Dr. I Jonathan Amster. I'm forever grateful for both your tutelage, kindness and giving me this eye-opening experience. It was more than I could ask for. The group has been great over the years, I hope these words will serve me well as a memory of everyone. The trips we've been on, the conversations we have, I wouldn't trade it for anything.

To my DSH family. Yes, family. You guys and girls made me feel like I had a home. You're the brothers and sisters I never had, and even in the toughest times your spirit and attitude and smiles lifted me up. I don't know where I'd be without you guys, what sort of dark hole I'd be looking down. Instead I look up at the sky, bright and radiant and making me feel whole. Thank you so much.

To those who gave me criticism. To those who pressured me to do better. To those who showed me my own faults and told me I could be better. Thank you.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION TO GLYCOSAMINOGLYCAN MASS SPECTROMETRY

1.1 AN INTRODUCTION TO GLYCAN BIOLOGY

Mass spectrometry has been fundamental in developing the fields of genomics, proteomics, lipidomics and metabolomics – allowing researchers to answer important biological questions with remarkable speed, accuracy and repeatability. While some biomolecules can be analyzed with great precision under a rapid analysis platform with user-friendly software for automated data interpretation, glycomics lags behind its other biomolecular counterparts in this regard. This is not to say that glycomics is less important; in fact, a visual cross section from electron microscopy of an endothelial cell surface shows the sheer density of carbohydrates that are present [1].

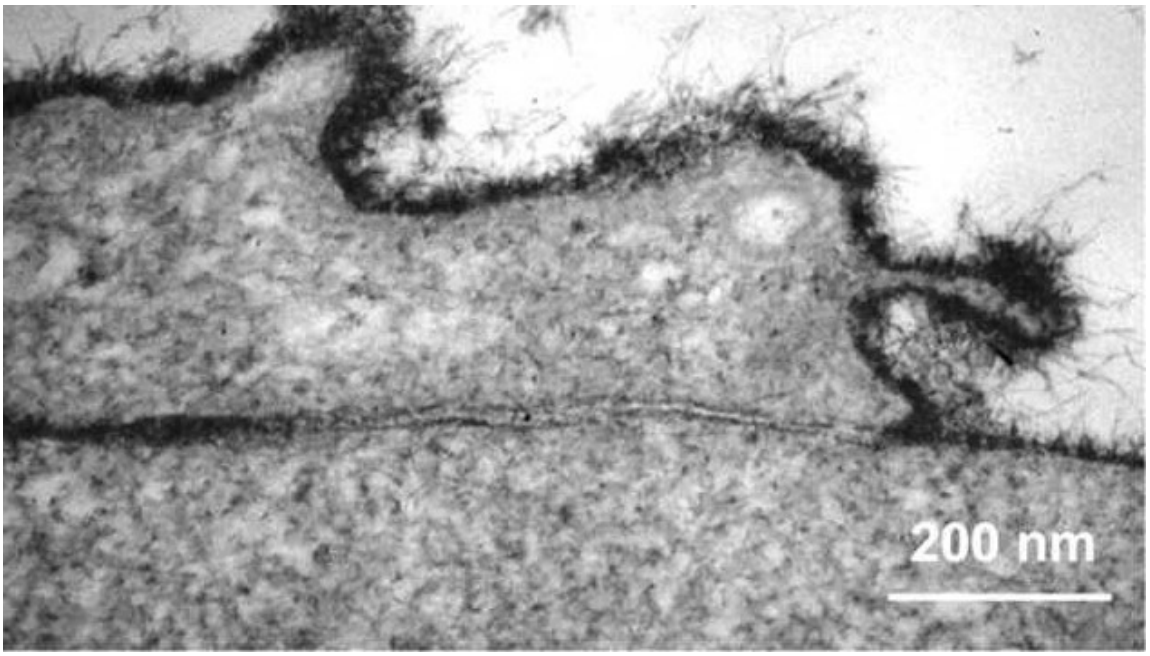


Figure 1.1. The glycocalyx of human umbilical vein endothelial cells. The surface contains a mixture of glycans and glycoconjugate molecules. A dense and highly complex mixture of biomolecules exist and participate in a multitude of biological activity. The direct relationship between these glycans and their effect on biological function still requires more research.

The cell surface is rich with carbohydrate-related molecules: polysaccharides, proteoglycans, glycoproteins and glycolipids are scattered throughout the surface. These molecules all serve a major biological role in some capacity, being a driving force for specific micro- or macromolecular interactions, and yet they are dramatically understudied. This is because ascertaining glycan structure has a multitude of difficulties: (1) identification and quantification of monosaccharides; (2) saccharide configuration (D- or L-); (3) branched vs unbranched; (4) glycan sequence; (5) α versus β anomers; (6) pyranose versus furanose rings; (7) position of the linkages (sugar rings can connect at different carbon positions to one another); (8) polymeric modifications / derivatizations (phosphate,

sulfate, acetate) that might be fundamental to glycan function; and (9) potential covalent linkages of glycan to a core protein.

The work and details discussed in the following section will focus on one type of glycan known as glycosaminoglycans (GAGs for short). GAGs are distinctly different from other glycan counterparts in that they are strictly linear carbohydrates consisting of a repeating uronic sugar (glucuronic acid/GlcA or iduronic acid/IdoA) and amino sugar (glucosamine/GlcN or galactosamine/GalN) copolymer. With established understanding of the polymeric backbone, GAGs avoid several problems associated with other forms of glycomics analysis. In mass spectrometry, the typical concerns for GAG structure are reduced to: sequence, polymeric modifications, D- vs L- configurations and covalent linkage to the protein. The last problem is one that has been circumvented with a few approaches prior to MS analysis, either by using GAGs that are produced enzymatically through treatment with a series of residue-specific cleavage enzymes[2] or GAGs synthesized in organic chemistry laboratories with high purity [3]. However, these methods would not work for studying native, intact GAGs, which are instead released from the proteoglycan core attached at the reducing end and then separated and fractionated extensively prior to mass spectrometry analysis [4-5].

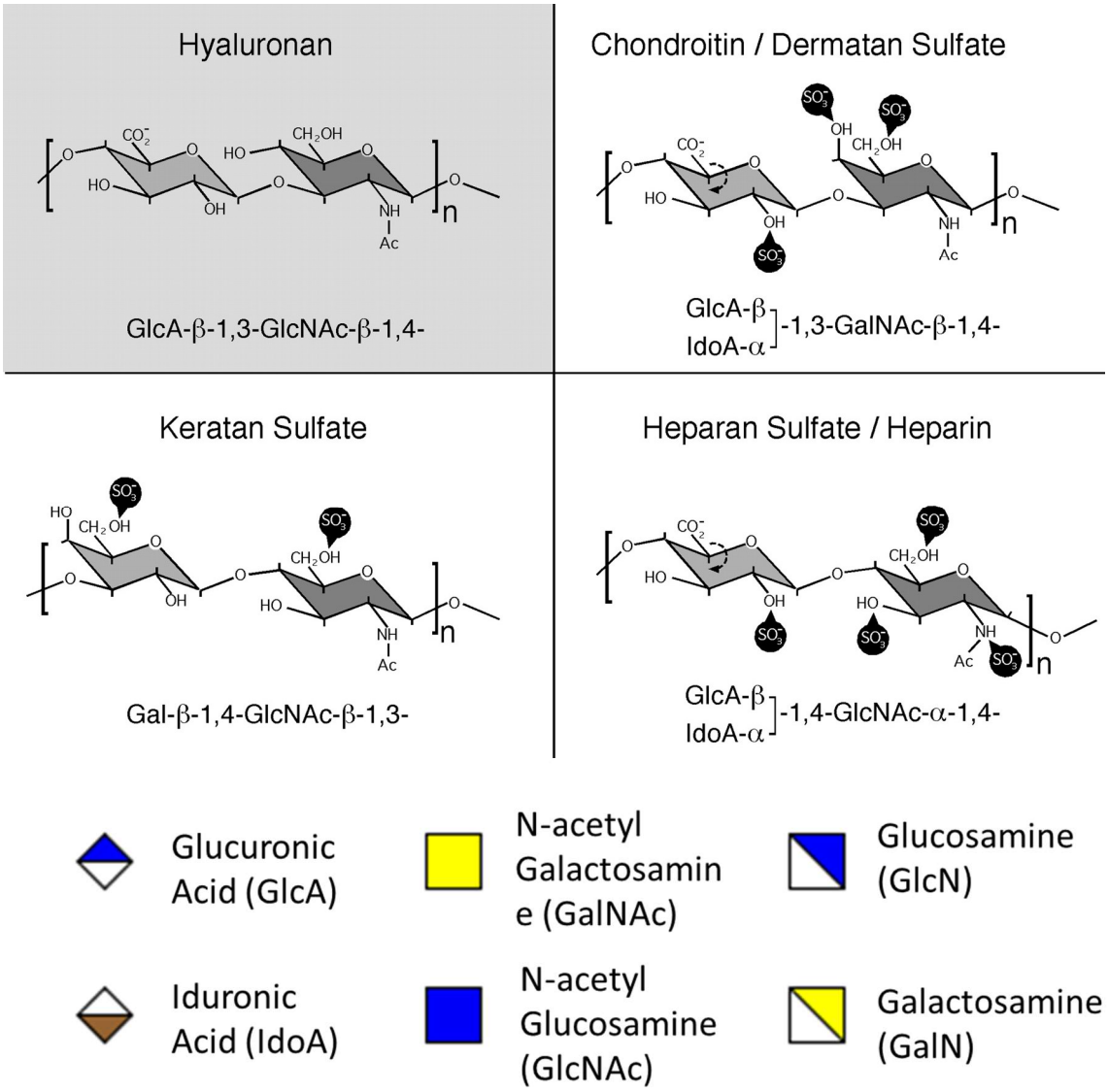


Figure 1.2. The four primary families of GAGs. Hyaluronan, keratan sulfate, chondroitin/dermatan sulfate and heparan sulfate/heparin each contain a repeating polymeric backbone. Black bubbles containing an 'SO_3' in the diagram indicate positions where a possible SO_3 modification could occur. Glycans are often shortened down to a cartoon representation as shown above.

GAGs are grouped into four primary families with their linkages well understood. Hyaluronan is the only GAG family not synthesized in the Golgi apparatus, instead being formed in the cellular plasma membrane from integral membrane synthases and

characterized by a polymeric backbone of β-D-glucuronic acid (β1-3 linkage) to a β-D-N-acetylglucosamine (β1-4 linkage). Hyaluronan exist within the extracellular matrix and is involved in physiological functions such as lubrication, water homeostasis, filtering effects and regulation of plasma protein distribution [6]. Hyaluronan's biological activities are numerous, with elevated concentrations linked to various types of cancer or oncogenic signaling pathways [7-11], being involved in arthritis development [12-13], and having regulatory effects in vascular disease and diabetes [14-15]. Hyaluronan is also the only non-sulfated GAG and one of the most extensively studied GAGs due to its simple structure and lack of any major microheterogeneities.

Structural microheterogeneity occurs within the other three families of GAGs: keratan sulfate (KS), chondroitin/dermatan sulfate (CS/DS), and heparin/heparan sulfate (Hp/HS). Nonetheless, sulfated GAGs are also critical in biology: KS is a hydrating and signaling agent in cornea and cartilage [16]; CS/DS has been found to be an effective treatment for osteoarthritis and maintaining tissue structural integrity [17-23], modulation of gut microbiomes [24], muscle development and axon guidance in zebrafish [25]; lastly, Hp/HS have played important roles in embro- and tumorigenesis in female reproductive systems [26], cellular activity in adipose derived stem cells [27], antimetastic activities of micelles [28] and blood coagulation by influencing protein binding of antithrombin III [29-31].

Microheterogeneities make sulfated GAGs significantly more difficult to characterize and are a result of the non-template driven biosynthetic pathway of KS, CS/DS and Hp/HS starting in the endoplasmic reticulum and propagating in the Golgi apparatus.

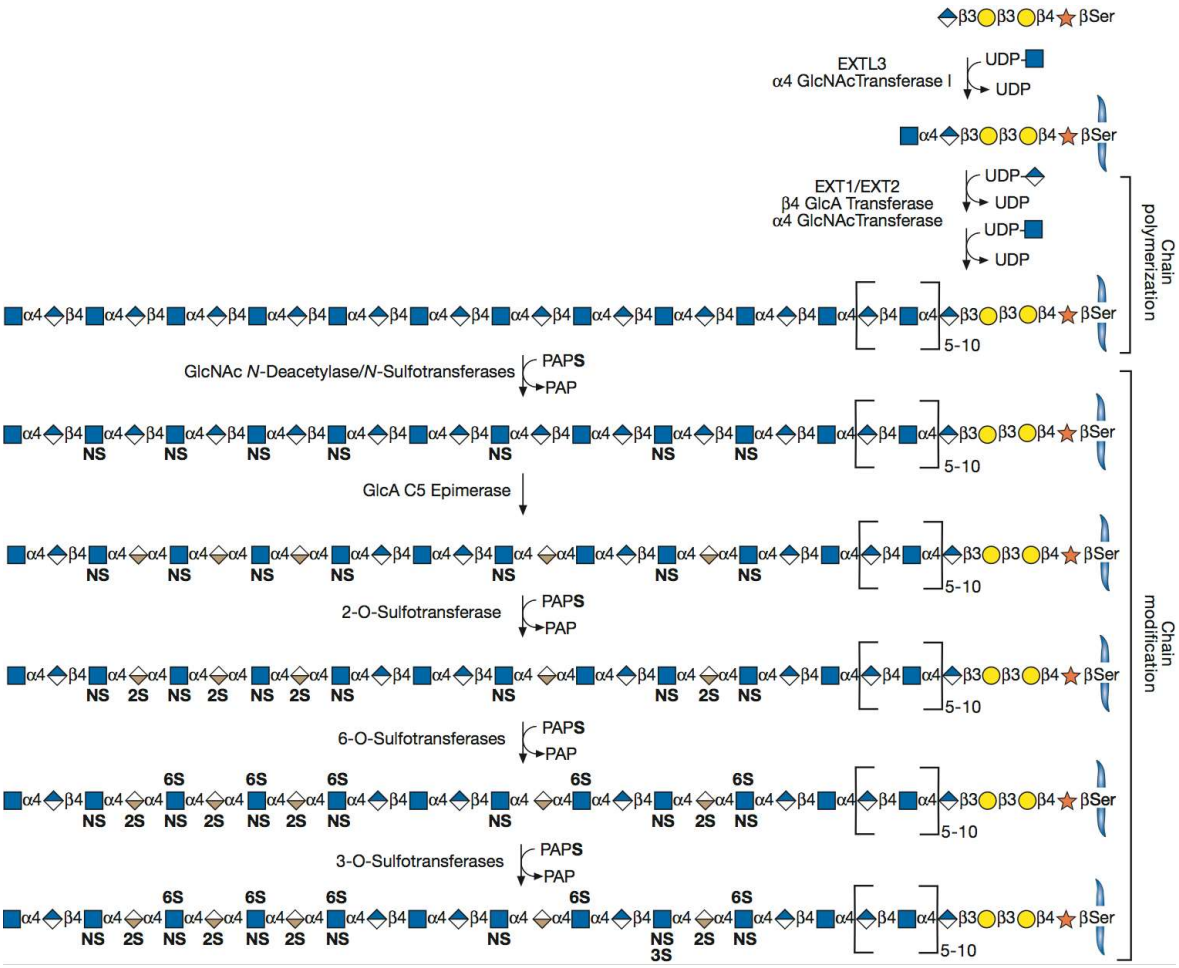


Figure 1.3. The chain polymerization of heparan sulfate / heparin GAGs within a proteoglycan. Alternating α4 GlcNAc transferase and β4 GlcA transferase propagate the GAG chain bound to a protein at a serine residue. Within the Golgi apparatus, possible chain modifications occur from multiple enzymes. Site specific sulfotransferases modify specific positions with an SO_3 group, epimerases change the C-5 uronic acid stereochemistry. N-deacetylases can also remove acetyl groups from amine sugars. Rarely do these transferase reactions go to completion, resulting in only partially modified regions within a chain.

The initial chain polymerization process of GAGs creates the repeating disaccharide backbone, but various enzymes can modify the chain within the Golgi. Often,

these enzymatic reactions do not go to completion, leaving parts of the chain unmodified in a way that is not predictable. Sulfotransferase enzymes promote the sulfation of specific positions on sugar residue and can occur with any sulfated GAG. For Hp/HS, deacetylase enzymes also combine with sulfotransferases to remove the acetyl groups from the amino sugar and replace it with the a sulfo-modification [32]. This higher complexity and seemingly random distribution of modifications is the most challenging aspect of GAG analysis using MS.

Nonetheless, for sulfated GAGs, structure is king. The modification patterns dictate GAG functionality and biological activity; some GAGs have been shown to have little to no effect on protein binding conformation if the modification patterns lack specific features [31, 33]. The goal of GAG MS should be to assign structure using optimal experimental conditions and a well-developed platform for easy and efficient interpretation.

1.2 GLYCOSAMINOGLYCAN ANALYSIS WITH MASS SPECTROMETRY

To date, glycomics is one of the less refined areas of biomolecular analysis in mass spectrometry. The subsection of this work that is related to GAGs is even more narrow and less developed. This section will discuss methods and strategies used commonly for GAG analysis with Fourier Transform Mass Spectrometry (FTMS). Determining GAG structure and characteristics from mass spectrometry requires expertise in understanding experimental ionization parameters and ion activation methods as well as methods for calculating composition and tools for interpreting fragments. GAG structural analysis is somewhat comparable to that of top-down proteomics: an MS^1 provides insight on the

overarching structural features while fragmentation of an isolated precursor with tandem mass spectrometry (MS^2/MS^n) can be used to determine sequence and locations of important modifications. Complete structural characterization is both dependent on sample complexity, such as highly sulfated versus lowly sulfated samples, and the number and quality of fragments that can provide fine structural detail.

1.2.1 Ionization.

Contrary to routine protein analysis, GAGs are normally analyzed using negative-ion mass spectrometry as the acidic uronic sugar of GAGs containing a carboxyl group is a likely candidate for deprotonation. Furthermore, certain GAG families (CS/DS and Hp/HS) contain multiple labile sulfo- modification which can be deprotonated as well. Efficient ionization is often a tricky process that requires low energy sampling methods and can be performed both with online separation as well as direct infusion. Pre-injection treatment of GAGs is not necessary but has been shown to be successful for structural analysis. A common process to promote a more uniform ionization of both acidic and basic sugar residues is methylation of the $-NH_2$, -OH, and -COOH moieties, exchanging the H atom with a methyl group [34]. This derivatization technique has been known to yield different tandem MS fragmentation patterns compared to that of underivatized GAGs but is not short of fragments capable of determining GAG structure [35-41].

To observe samples in MS, low energy ionization techniques are required to minimize fragmentation and SO_3 loss prior to being detected. Matrix Assisted Laser Desorption (MALDI) in which a dried sample spot is irradiated by ultraviolet frequencies has been used for glycan ionization [42]. Matrix optimization has been studied extensively

for carbohydrates but is dependent on sample and in some cases concentration. Matrices of substituted benzoic acids, 3-aminoquinoline, mercaptobenzothiazoles, beta-carbolines, osazones, ferulic acid and hydroxyacetophenones have all shown varying degrees of success for glycan ionization [42]. MALDI has demonstrated a higher sensitivity towards glycans and ionizes at higher mass ranges compared to electrospray ionization (ESI). However, a major caveat occurs in the relative simplicity of MALDI spectra. Single charged ions are often formed in MALDI and provide less structurally significant peaks compared to ESI, not to mention the tendency for more in-source fragmentation due to the higher energetics of laser desorption techniques.

ESI of GAGs is performed by passing a mixture of dilute GAG solution (~0.001 – 0.1 mg/mL) through a thin diameter syringe, needle or tip with a potential difference of ~1-4kV between the source and capillary of a mass spectrometer. The formation of charged droplets occurs due both potential difference and high temperature, which is then vaporized via a drying gas before entering the instrument [43]. A key advantage of using ESI over MALDI is the ability to consistently generate ions of multiple charge, making it more feasible for ion activation at the MS^n level and therefore more structurally informative spectra. Furthermore, the ability to couple ESI to separation techniques such as liquid chromatography (LC) [44-46], ion-mobility (IM) [31, 33] and capillary electrophoresis (CE) has made it an excellent option for high-throughput analysis of mixtures.

1.2.2 Structural Characterization.

The first step in determining glycan structure is acquisition of a high resolution MS^1 with low mass error (typically <15ppm, although often <5ppm). High mass accuracy is desired as it allows for unambiguous determination of GAG composition: degree of polymerization, number of SO_3 and number of acetyl groups can be assigned by mathematical calculation. A standard approach to matching composition is using a table of mass values for all desired chain lengths and numbers of modifications. However, as mass accuracy decreases, peak(s) in the MS^1 are more likely to match multiple compositions and thus reduces the ability to determine structure based on tandem mass spectrum (MS^2) fragmentation.

Once a composition is determined, more structural information regarding the location of modifications can be determined based on MS^2. Nomenclature for fragments that result from ion activation techniques has been established by Domon and Costello [47]:

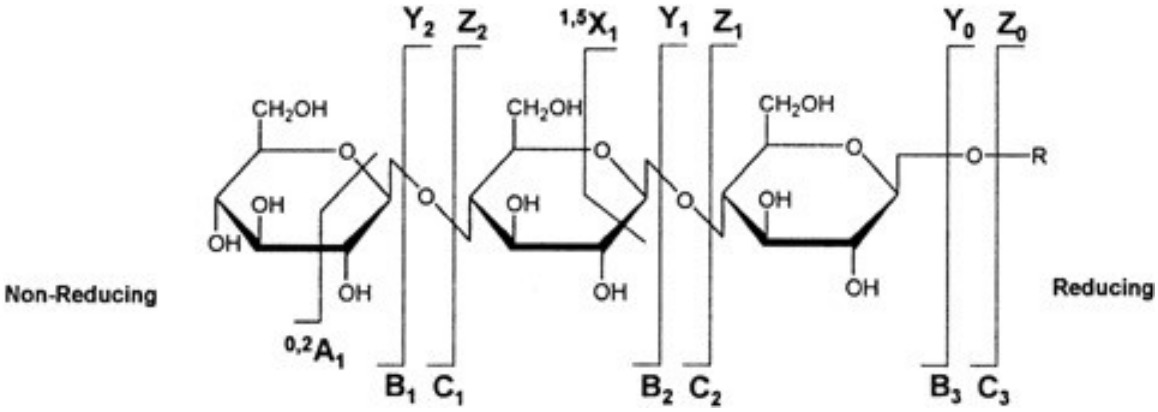


Figure 1.4. The standard Domon-Costello nomenclature for fragmentation of GAGs. Capital letters (A,B,C, X, Y, Z) are used for glycans to differentiate from lower case letters commonly used for protein MS^n analysis.

B, C, Y and Z fragments are referred to as glycosidic and occur between sugar residues. Cleavages derived from the non-reducing end prior to an oxygen are regarded as B fragments, whereas C fragments occur after the oxygen. From the reducing end, Y fragments are complementary to B and cleave the same bond; as is also the case with Z and C fragments. A subscript is used following the letter to indicate which residue the cleavage occurs on. Cleavages that occur within a ring are referred to as cross-ring fragmentations and are denoted with a superscript of two numbers prior to the A or X. The superscript is indicative of the bond positions on the sugar residue that are cleaved, with a position of 0 being the bond occurring immediately right of in-ring oxygen atom. Cross ring A fragments from the non-reducing end contain a subscript for the residue that they stem from, while X fragments start with a subscript of 0 rather than 1. In this manner, the subscript combination of A and X fragments on the same residue will sum to the residue where it is located.

The presence of fragments is essential for determining structure. An individual glycosidic cleavage at some arbitrary point along a GAG dictates the number of residues, what type of residues and how many modifications are on either side of that fragment. A singular glycosidic fragment does little more than provide this simple detail but a series of complementary fragments can be used to roughly sequence the number of modifications on each residue. For example, a B_1, B_2 and B_3 fragment series with masses matching that of HexA, GlcA+GalN+SO_3, 2GlcA+GalN+SO_3 (HexA = hexuronic acid, GalN = galactosamine) reveals that the non-reducing end of the sequence starts with a glucuronic acid and has an SO_3 modification on the galactosamine residue.

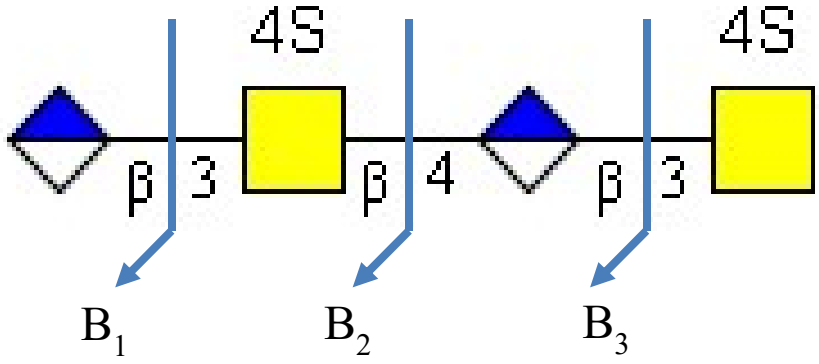


Figure 1.5. A chondroitin sulfate tetrasaccharide can be sequenced with a series of 3 glycosidic cleavages. The mass difference between two glycosidic fragments adjacent to a residue can be used to determine whether 1 or more SO_3 groups are present on that residue.

Determining the repeating disaccharide unit and approximately where modifications are located is the extent of information that can be meaningfully gained from glycosidic fragments alone. Glycosidic fragments can also be isobaric depending on the symmetry of the structure being analyzed: B and Z fragment ions will be the same mass when cleaving certain bonds if they produce fragments containing the same number of residues and modifications and are indistinguishable from one another in the spectrum.

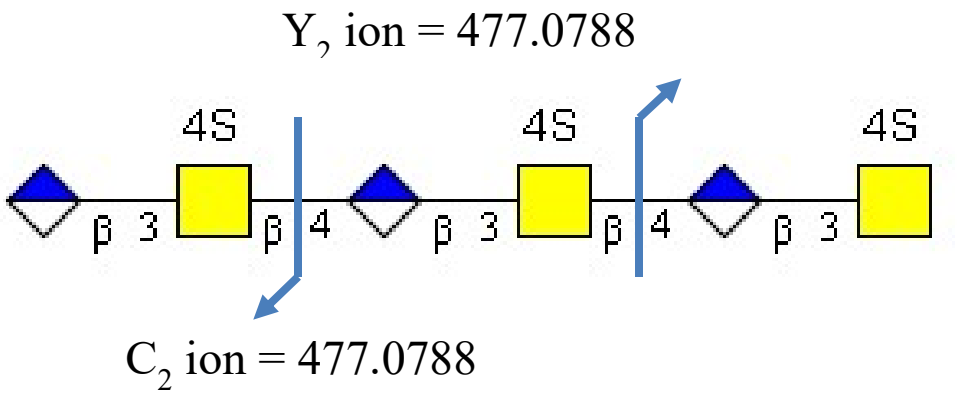


Figure 1.6. A common problem that occurs when analyzing GAGs that have been enzymatically produced is the possibility of isobaric ions. Here, the C_2 and Y_2 fragment ions are the same mass and make it difficult to determine structure without the presence of other fragments. The reducing end can be modified with additional mass tags, which will increase the mass of the Y_2 ions and make it possible to differentiate these fragments.

GAG structural analysis is often complicated by isobaric fragments and various steps such as derivatization [48-49] or synthetically produced GAGs with a modified reducing end [3] avoid this problem entirely. Moreover, GAGs released in their intact form from proteoglycans contain a linker section with non-homogenous sugar subunits and serve as a unique mass tag [4-5].

Cross-ring fragments cleaving specific bonds of the rings of sugar residues may yield more information regarding the specific position of modifications. For example, a $^{0,2}A$ or $^{0,2}X$ cleavage will effectively isolate the 2nd carbon atom on the residue that is being cleaved – in combination with glycosidic fragments related to that residue, information regarding what sort of modification exists at the 2-position is possible.

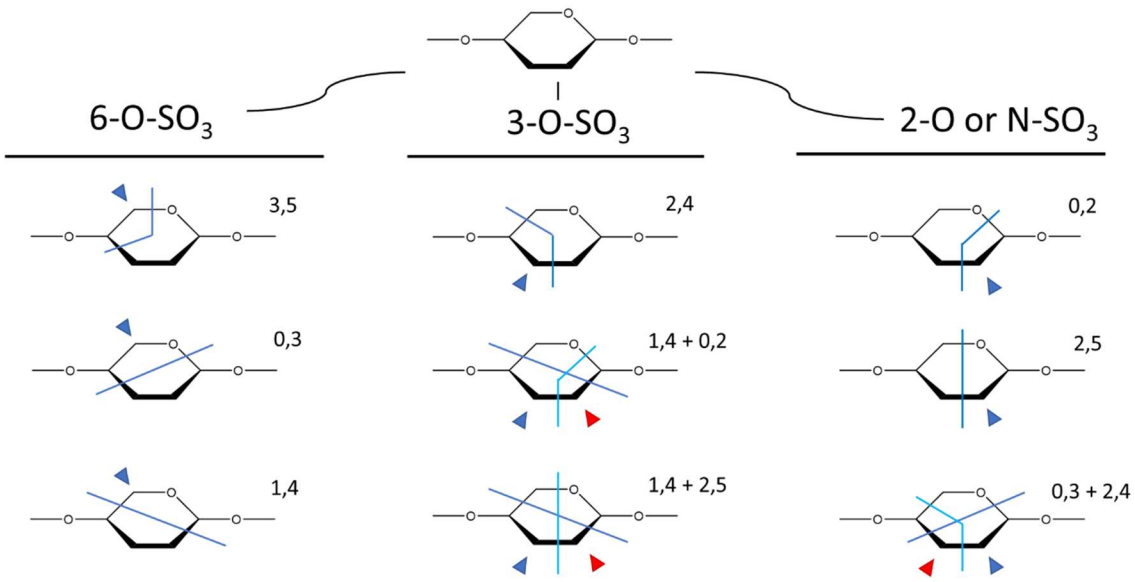


Figure 1.7. A schematic for modification positions that can be determined with cross-ring fragments. These fragments are typically combined with a glycosidic fragment to determine if a modification is present. Blue arrows point to fragments that can be determined with a single cross-ring fragment, while red arrows indicate positions determined using 2 specific cross-ring fragments.

The combination of cross-ring fragments and glycosidic fragments is useful when more than one modification can appear on a residue. However, for the uronic sugar residue of GAGs, only one sulfo- modification is possible. In this specific instance, only the 2-O position of the uronic sugar can be modified. A pair of glycosidic fragments would be sufficient to determine if the residue is modified; any cross-ring fragments would only further verify what is already assumed to be true. While the sulfo- modification can be determined easily, the stereochemistry of the C-5 uronic sugar is more difficult to access through mass spectrometry. Characterization of the epimeric center is dependent on the observation of diagnostic fragment ions and is dependent on ion-activation method. Primarily, electron activation methods such as electron detachment dissociation (EDD)

have yielded fragments that are exclusive to glucuronic acid [50]. More exhaustive studies of synthetic tetrasaccharide heparin sulfate samples also suggest the possibility of diagnostic ratios for determining the presence of glucuronic versus iduronic acid [51]. This point will be explored further in a following section.

The final point of interest regarding ion activation is the lability of sulfo-modifications and their tendency for a hydrogen rearrangement reaction in which the SO_3 group is lost prior to fragmentation [52]:

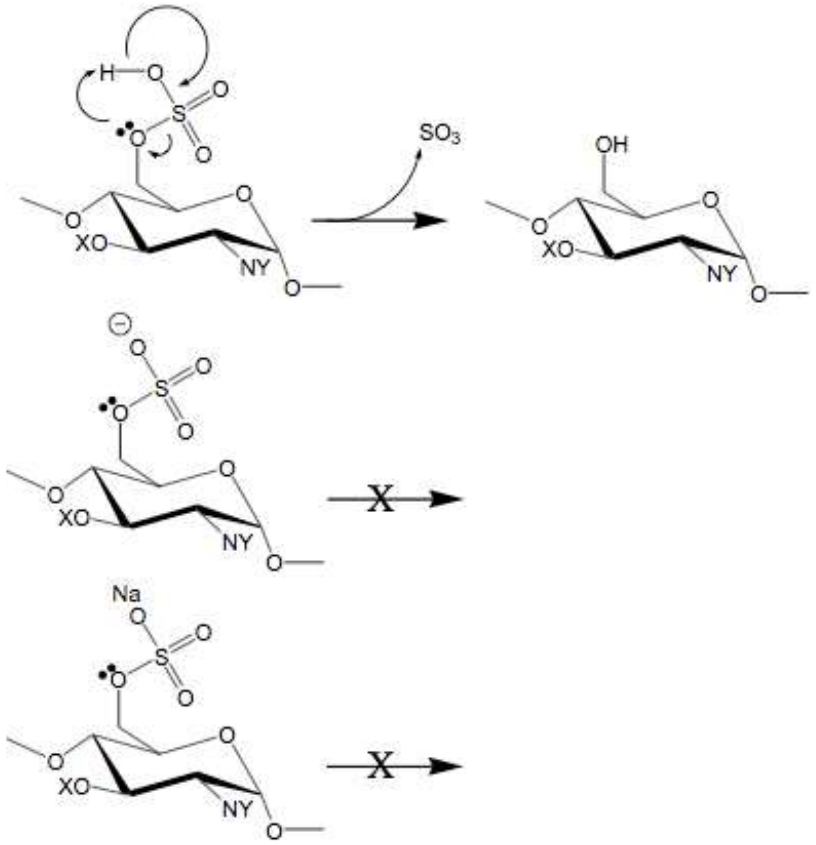


Figure 1.8. Hydrogen rearrangement can occur with the sulfate half-ester bond unless the SO_3 group is either A) deprotonated or B) adducted to a metal cation such as sodium (Na^+).

Loss of the SO_3 group prior to any fragmentation hinders structural characterization and sequencing: positional information is lost alongside the SO_3 group. Complete deprotonation of all ionizable SO_3 groups is the best solution for minimizing modification loss without additional pre-injection sample treatments such as derivatization. In the mass spectrometer, the precursor ion with a charge equal or greater than the number of SO_3 groups should be selection for ion activation. Alternatively, the addition of metal adducts can serve a similar role in preventing hydrogen rearrangement. Addition of sodium hydroxide (NaOH) is a proven method for promoting Na-H exchange, leading to increased stability of the sulfo-modification [53].

1.2.3 Ion Activation and Fragmentation Techniques.

Collision Activation. GAGs have been fragmented in both FT-ICR and Orbitrap instruments using common collision-based threshold methods that are routine to most MS platforms. CID [54] (collision induced dissociation) and HCD [55] (higher energy collision dissociation) have been used to obtain meaningful fragmentation patterns that provide insight on GAG structure. Sequence informative glycosidic cleavage are often the most commonly observed and most intense fragments when using collision-based ion activation. However, CID also tends to yield fragments containing the loss of the sulfo-modification which confounds structural determination. Cross-ring fragments can also be observed but are few and far between compared to other ion activation methods discussed later. Nonetheless, structural characterization of both pure standards and unknown structures has been realized. Low molecular weight heparin-based drug Arixtra has been characterized using CID in FT-ICR with the assistance of Na^+/H^+ exchange to stabilize against SO_3 loss

[53]. CID has also been shown to have potentially diagnostic fragment ions that differentiate between glucuronic and iduronic acid for choindroitin and dermatan sulfates from 4 to 10 degrees of polymerization [56]. CID fragmentation has been used to characterize intact bikunin choindrotin sulfate GAGs released from its proteoglycan using *de novo* manual interpretation coupled with disaccharide analysis [4]. Even without an abundance of cross-ring fragmentation, a high density of sequence informative glycosidic fragments can be interpreted to generate a complete structure.

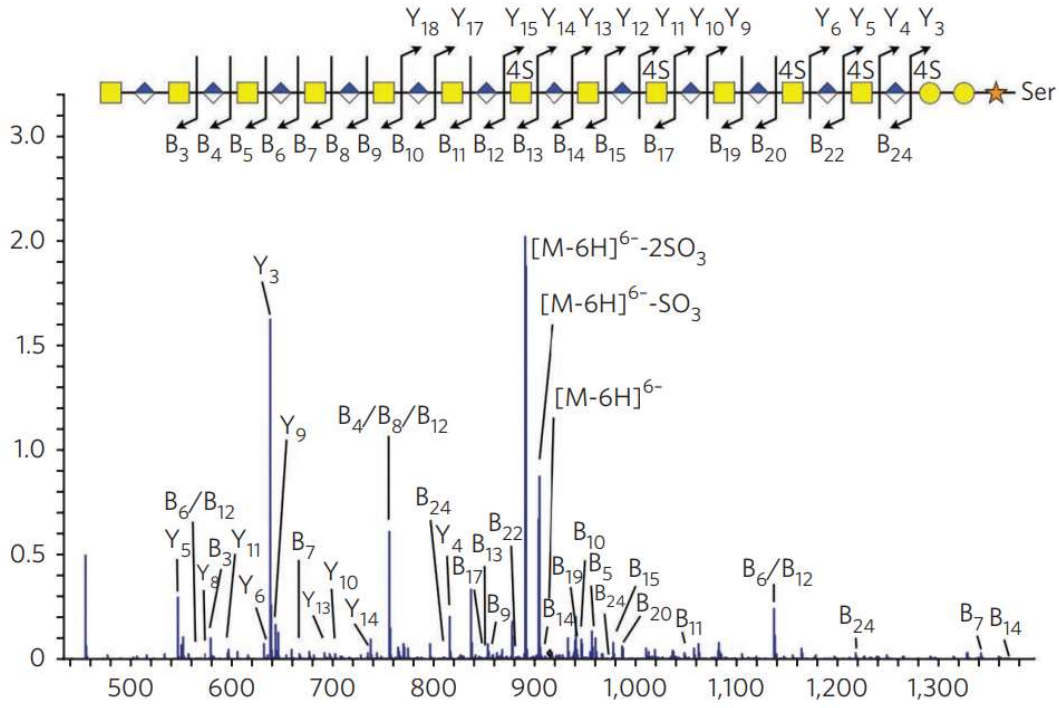


Figure 1.9. A CID-FT-ICR-MS^2 spectrum of bikunin GAG. Plentiful amounts of glycosidic fragmentation (B and Y) allows determination of where SO_3 groups (denoted 'S' in figure) are located within the chain. Disaccharide compositional analysis was used to determine that only the 4-O-position of GalNAcs are modified.

Electron Activation. Unambiguous determination of site-specific modification position requires fragmentation of the GAG backbone. Modern electron-based methods

have become a popular option for this purpose. Electron activation involves a multiply charged molecular ion transferring an electron to the analyte of interest, generating a radical ion that undergoes fragmentation [57-58]. Electron activation favors the production of both glycosidic and cross-ring fragmentation and more importantly retains information regarding sulfo-modification position. This technique is especially attractive since it avoids the need for chemical derivatization and/or metal adduction that are typically associated with stabilizing the sulfate group, thus reducing complexity of the overall experimental workflow.

A promising and popular FT-ICR technique for GAG analysis is electron detachment dissociation [59-60] (EDD) as it yields a high volume of sequence informative glycosidic and cross-ring fragments as well as provides potentially diagnostic fragments for determination of uronic sugar stereochemistry. An abundance of cross ring fragmentation in highly sulfated GAG compounds allows for site-specific determination of modifications.

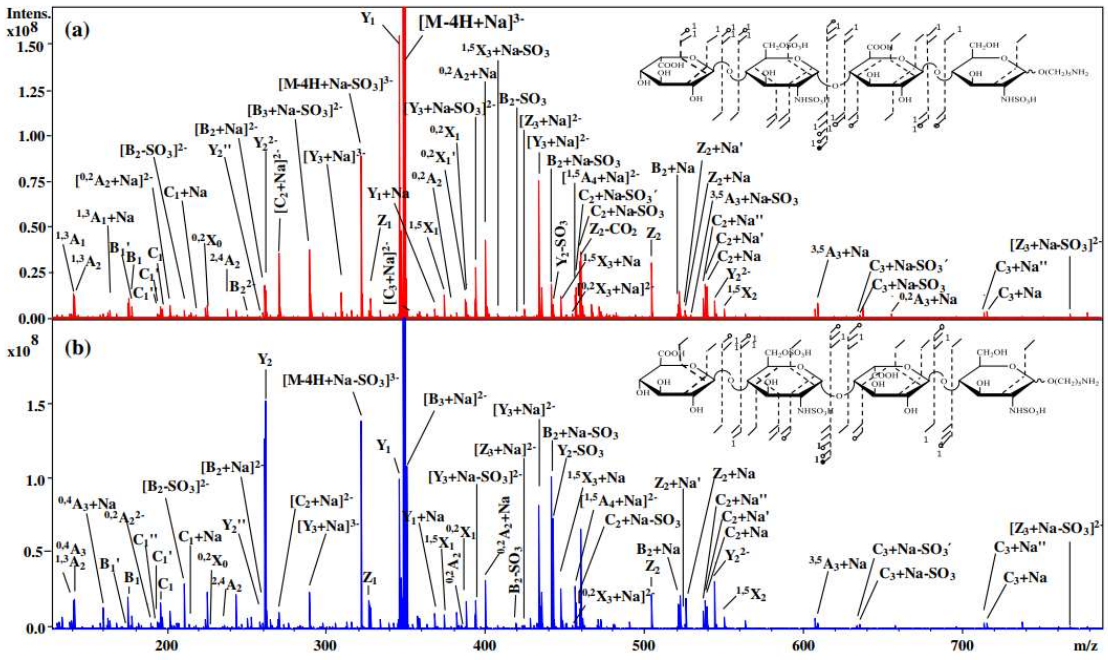


Figure 1.10. Electron detachment dissociation (EDD) of two tetrasaccharide samples [61]. The two structures differ by only their C-5 uronic sugar stereochemistry on the reducing end uronic sugar residue. An abundance of cross ring and glycosidic fragmentation allows for more complete structural determination.

Electron detachment dissociation studies on moderately sulfated tetrasaccharides suggests that only glucuronic acid residues contain the ^{0,2}A, B_3-H, and B_3-HCO_2 fragment ions and such fragments are absent when iduronic acid is present [50]. Moreover, exhaustive studies of heparin sulfate tetrasaccharides of varying degrees of sulfo modification (0.5-2.5 sulfates per dissacharide) have yielded a diagnostic ratio based on fragment intensity for distinguishing glucuronic acid vs iduronic acid [61]:

$$Diagnostic\ Ratio = log\left(\frac{1\sum(B_3, Y_1, C_2, Z_2}{3\sum(Y_2, {}^{1,5}X_2)}\right)$$

The diagnostic ratio value is positive for glucuronic acid and negative for iduronic acid and remains true for all degrees of sulfo modification on synthetic tetrasaccharides. The question remains as to whether the ratio continues to be diagnostic for structures larger than tetrasaccharides and is currently under investigation.

While electron detachment provides structurally informative fragments in high abundance, it is a relatively slow technique intended for FT-ICR and is not conducive to a high throughput platform. Long acquisition times with EDD becomes a temporal bottleneck and are not tractable with online separation time scales. Negative electron transfer dissociation (NETD) has shown promise as a technique for characterizing glycosaminoglycans [62-64] and can be used within an Orbitrap MS with a significantly improved throughput rate (1s for EDD vs 0.1s for NETD).

NETD of GAGs involves the use of xenon or fluoranthene as a reagent gas, which are radical anions produced in a chemical ionization (CI) source external to the ion trap in which the NETD reaction occurs. A multiply charged ion precursor reactions with the radical anions and an electron is transferred from the radical anion to the precursor, generation an odd-electron species that undergoes fragmentation.

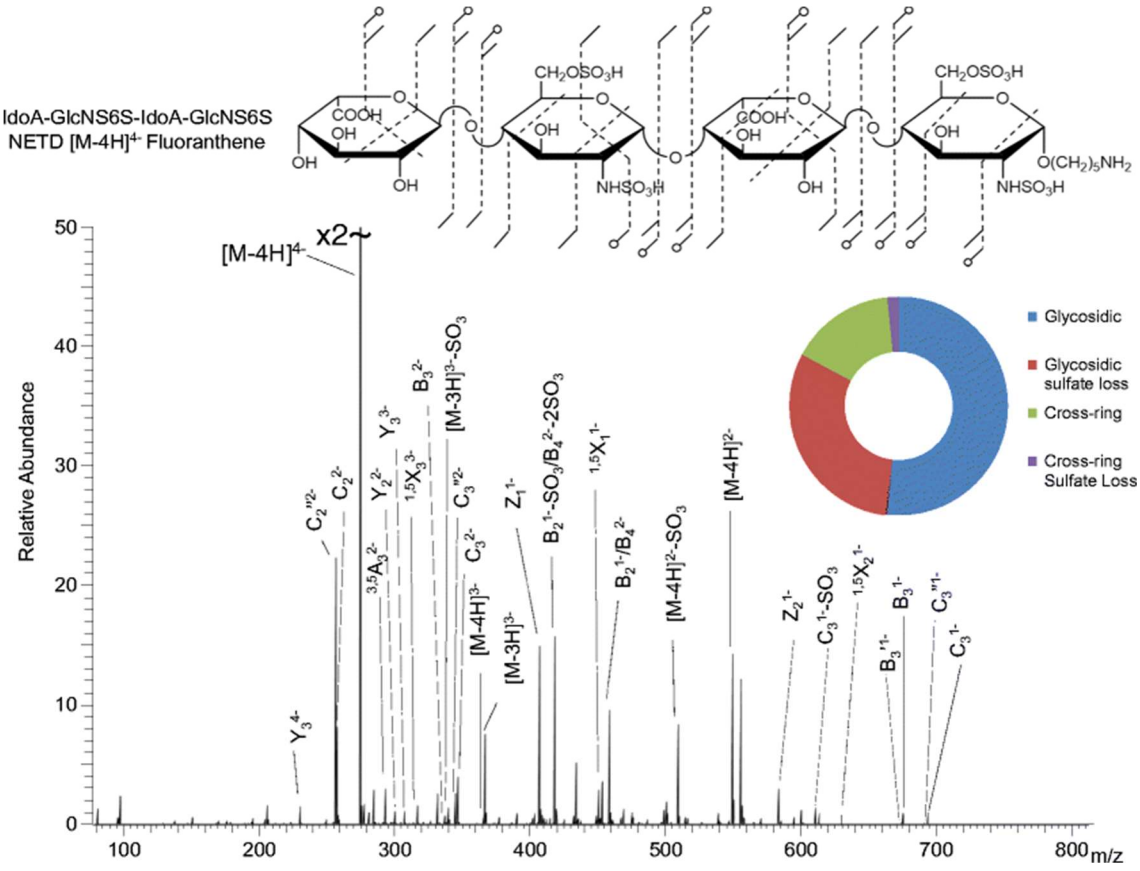


Figure 1.11. Synthetic heparin tetrasaccharide standard that was fragmented using NETD [63]. NETD provides a comparable level of sequence informative fragmentation to that of EDD but works on a more rapid timescale suitable for online separation.

Photodissociation. The recent development of ultraviolet photodissociation (UVPD) as an ion activation method for proteomics [65-66] has paved the way for applications in other biomolecules in both ICR and Orbitrap instruments. Glycans are no exception [67] and have been fragmented using a 193 nm laser source. Photodissociation has been shown to produce fragments similar to electron-based methods [68], leading to the formation of cross-ring fragments that can be used as diagnostic indicators for modification positions. UVPD is comparable to NETD in terms of both fragmentation abundance and diversity and also works well on an Orbitrap timescale. Other UVPD parameters such as excitation

wavelength (193 nm vs 213 nm), high pressure versus low pressure ion activation and fragmentation efficiency should be investigated more thoroughly before being declared as a superior alternative to NETD.

1.2.4 Separation and Ion Mobility.

Glycosaminoglycans of a defined composition contain numerous structural isomers and scales exponentially as the length of the GAG increases. MS^2 can provide meaningful insight into structure when a singular component is being examined. However, in the case of a mixture it becomes far more difficult to assign structure due to structural isomerism. Moreover, determination of the C-5 uronic sugar stereochemistry by fragmentation in the MS^2 alone is highly specific to sample type and ion activation method. Modern separation methods including liquid chromatography (LC), (sometimes coupled with hydrophilic interaction chromatography (HILIC) or porous graphitized carbon (PGC) stationary phases) and capillary electrophoresis (CE) separations have all been used for GAG separation and characterization. Ion mobility (IM) methods have also been used to investigate GAGs, although to date no consensus exists for the most optimal GAG separation technique.

Liquid Chromatography (LC). Perhaps the most popular method of separation for biomolecules, LC has been a method for successfully separating GAGs with reversed phase (RP) chromatography in which a polar mobile phase elutes through a hydrophobic stationary phase, with C_{18} being the preferred column of choice. Chemical derivatization and permethylation has been the prominent pre-separation protocol [69] and coupled with

LC-MS^n, can be used as a method for characterizing GAG mixtures, able to distinguish CS/DS isomers [49] and structurally diverse Hp/HS GAGs [62]. Alternative derivatization strategies using 2-aminoacridone (AMAC) have been applied to successfully differentiate pharmaceutical and low molecular weight Hp/HS [48]. LC is also an excellent tool for analyzing disaccharide compositions derived from digestion native GAGs of proteoglycans using various GAG-lyases [70], which can be coupled to direct infusion of full length GAGs to simplify spectra interpretation and structure elucidation [4-5].

Ion-pairing reagents can be added to the mobile phase of RP-LC (RPIP-LC) to enhance the separation of Hp/HS digests. Reagents such as tributylamine (TrBA), dibutylamine (DBA) and pentylamine(PTA) are lipophilic reagents that work well with the mobile phase of acetonitrile and methanol for characterizing pharmaceutical heparin [71] and Hp/HS disaccharides [72-73].

A microfluidic chip-based LC-MS platform has also been successful for the compositional profiling of GAGs from tissue samples [74]. Amide hydrophilic interaction chromatography (amide-HILIC) is used in conjunction with a microfluidic chip LC-MS to yield efficient and robust separation of GAGs and other acidic glycans while also reducing background from other biological contaminants using trapping cartridges. The method is also useful for CS/DS and Hp/HS oligosaccharide analysis and has been shown to welcome the addition of metal cations or supercharging reagents (e.g. sulfolane) to reduce SO_3 decomposition during CID [75]. Porous graphitized carbon (PGC) can serve as a stationary phase for GAG analysis in LC-MS. PGC-LC-MS has been shown to be sensitive to nanomolar concentrations of HA, KS and HS from synovial fluid samples digested by

hyaluronidase, keratanase and heparinase respectively [76]. PGC-LC-MS workflows have also been used to quantify Hp/HS nitrous acid depolymeriation products [77].

Capillary Electrophoresis (CE). A most recently innovation in the world of GAG separation has been reverse polarity capillary zone electrophoresis (CZE) coupled to negative ion mode MS. Separation of compounds in CE depends on the differential migration time of samples in an applied electric field. For negative mode, the inner wall of the separation capillary is cation-coated with either N-(6-aminohexyl) aminomethyltriethoxysilane (AHS) or dichlorodimethylsilane (DMS) and coupled to a sheath flow interface to allow for rapid and reproducible migration of GAGs to an Orbitrap MS. CE-MS has been able to separate Hp/HS disaccharide standards [78], heparin tetrasaccharides [79] previously analyzed using LC-MS [80], synthetic tetrasaccharide standards [3], and pharmaceutical Enoxaparin. CE remains a relative new field in GAG separation but has been shown to be able to separate GAGs that differ in mass, are structural isomers, are epimers, or even in some cases anomers [79].

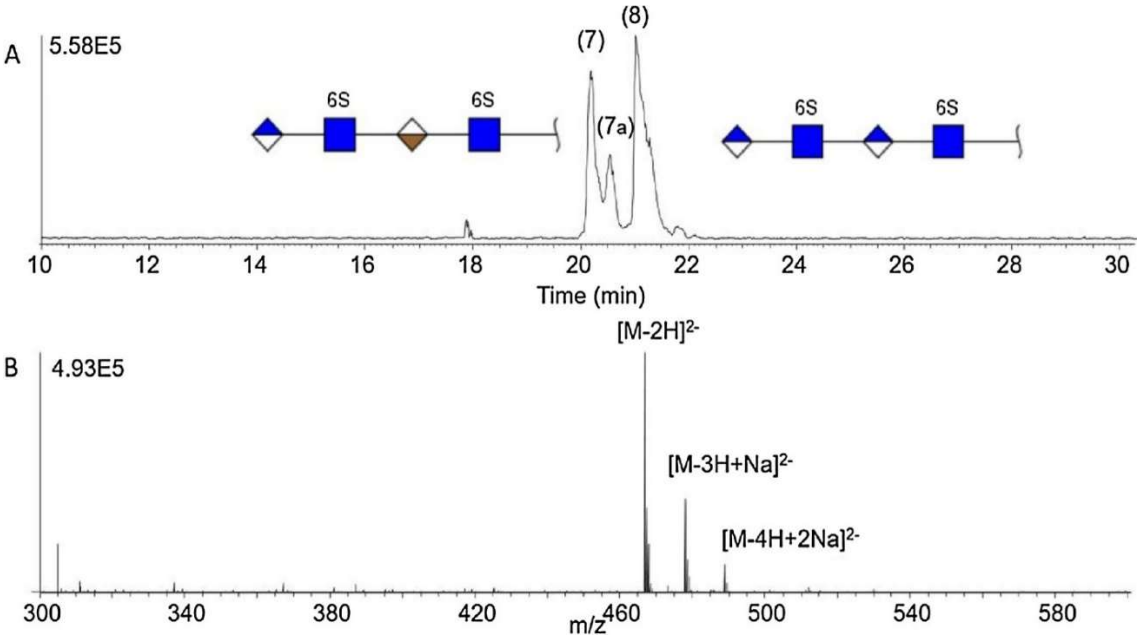


Figure 1.12. The CZE separation of two structures with the exact same mass and modification pattern. Different migration times occur based solely on the epimeric center of the reducing end uronic acid. Minor peaks within the electropherogram are anomeric components.

Ion Mobility (IM). Ion mobility separation occurs based on a molecule's interaction in the gas-phase with a buffer gas in either a varying or fixed electric potential through a sample drift region. The combination of mass spectrometry and IM provides information both on mass and shape (in a measurement known as the collision cross section). IM can therefore be used to not only separate GAG isomers but also examine structural changes in binding characteristics of GAGs with other molecules.

Differential mobility / high-field asymmetric waveform ion mobility (FAIMS) applies an asymmetric and periodic waveform along the drift region of the separation device with an alternating high and low electric field using two electrodes. Ions are dispersed in a manner where they end up moving towards one electrode – a compensation

voltage (CV) can then be applied to compensate for ions of a particular mobility, allowing it to reach the mass spectrometer. FAIMS coupled to ESI-ICR experiments have been able to differentiate chondroitin sulfates of different lengths as well as heparan sulfate tetrasaccharide epimers differing only by their C-5 uronic sugar stereochemistry [81].

Traveling wave ion mobility (TWIMS) separates ions through a buffer gas using a dynamic electric field and can determine shaped-related characteristics through the measurement of collision cross sections (CCS) based on drift times. GAG separation based on structural conformations and characterization using MS^n with TWIMS is possible for HS octasaccharides [82]. Protein-GAG binding complexes and changes in structure have also been investigated. The binding affinity of pharmaceutical Arixtra (a heparin-like GAG) and similar heparin structures to Antithrombin III (ATIII) has been investigated using TWIMS. Changes in CCS of protein-GAG complex show preferential binding of ATIII to GAGs containing the 3-O-sulfation modification and is negligibly affected by the presence or absence of other modification groups [31]. Moreover, monitoring of CCS measurements of Fibroblast growth factors (FGFs) and their interaction with different structural variations of HS oligosaccharides has shown that FGF have higher binding affinities with specific GAG lengths up to 12 degrees of polymerization but lacked higher specificity for longer chains. Potential ideal protein:GAG (FGF:HS) binding ratios were found to be 2:1 and 3:1 [33].

IM is still a new development and its applications in the realm of GAG analysis has yet to be fully explored. The techniques, especially those using CCS measurements, show promise not only as a means of GAG characterization but also an avenue for exploring GAG binding interactions with other biomolecules.

1.3 AUTOMATED STRUCTURAL INTERPRETATION

1.3.1 Challenges in Automation.

The field of proteomics is filled with a plethora of user friendly automated analysis software packages. Generic and highly specific packages are available to the public; whether it be top-down or bottom-up proteomics, successful software often relies on heuristics based on the template-driven biosynthesis of proteins. The predictability of amino acid sequences lends itself to a more database-oriented method for rapid analysis. Automated interpretation and sequencing of GAGs, whose biosynthetic pathway is non-template driven, proves to be far too different from proteins for a similar framework of analysis. The current standard for glycan structural analysis is a software package known as GlycoWorkBench [83], which allows the user to draw a glycan structure (branched or linear) and produce the theoretical fragments from said structure. The fundamental caveat here is that structure must already be known, which is unlike any proteomic software counterparts which can determine protein and sequence without user supervision.

Database building for such a large scale and unpredictable set of biomolecules (1) is tedious: the shortest chain length that can be analyzed is 2 degrees of polymerization but there is no theoretical maximum length; meaning a well-rounded and universally useful databases needs to encapsulate a limitless number of possibilities. And (2): can yield high rates of false positives due to the homogenous nature of the GAG backbone coupled to a non-zero likelihood of SO_3 decomposition within the mass spectrum. The likelihood that a sulfate decomposition peak is matched against an incorrect structure is highly likely when database searching and additional database-searching criteria would have to be implemented to mitigate the effects of false positive matches.

A shift in the paradigm of analysis is necessary: *de novo* sequence has shown promise in proteomics as a means of determining peptide sequences based on mass differences of fragment peaks in MS^2 without the need for database comparison [84-85]. GAGs are excellent candidates for *de novo* approaches as their repeating polymeric backbone (uronic sugar and amino sugar) is consistent regardless of chain length and number of modifications. This allows an interpreter to specifically look for mass spacings between fragment peaks that correspond to the mass of either a uronic or amino sugar residue. Within this framework, one or more modifications (SO_3 and N-Acetylation) can also be considered at each residue. Manual interpretation of mass spectra using *de novo* methods can only be successful up to a practical limit for human interpretation. The number of possible combinations for a particular GAG scales exponentially as 2^n, where n is the number of sites of possible modification. Assuming an accurate mass MS^1 can be acquired and composition of the GAG can be assigned (i.e. GAG degree of polymerization, number of SO_3 groups, number of acetylation), the number of permutations is equal to *n-choose-k,* where n is the number of occupiable sites of modification and k is the number of modifications:

$$n - choose - k: \quad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

The sheer volume of possibilities begins to be problematic for GAGs beyond 6 degrees of polymerization but realistically impractical at 10-12 units depending on number of modifications, unless additional information regarding the sample such as disaccharide composition analysis is available.

1.3.2 Automated Software Suites.

Many analytical laboratories have developed methods, scripts and workflows for automated or semi-automated interpretation of GAG spectra. This section discusses each piece of software currently in press.

HOST (heparin/HS oligosaccharide sequencing tool) [86] is a computational tool designed for sequencing heparin/HS oligosaccharides using enzymatic digestion combined with ESI-MS^n. The method scores and returns the best matching sequences of GAGs based on disaccharide composition analysis, yielding predicted compositions and calculating expected fragmentation patterns *in silico*. Comparisons of theoretical fragments can then be compared to fragmentation of heparin/HS oligosaccharide MS^n data and is scored to return the most likely sequence. However, disaccharide analysis requires complete enzymatic digestion of the GAG using heparin lyase I, II and III over multiple hours of incubation (16 h), limiting the method's overall speed and applicability in a high-throughput GAG analysis platform.

Another piece of software known as GAG-ID [87] has been shown to discriminate and identify 21 synthetic tetrasaccharides eluted from LC-MS/MS using a scoring system based on peak intensities. It is the first of its kind to automated the interpretation of mixtures when coupled to LC-MS/MS but require complete chemical derivatization of the GAG by replacing all labile sulfate modifications with more stable acetyl groups. Much like HOST, derivatization may not be a viable option for universal GAG analysis.

HS-SEQ [88] is a *de novo* GAG sequencing computation framework that has been used to automate the structural identification of HS of dp4, 5, 6, 8 and 15. The method determines a precursor sequence (unmodified GAG backbone) and uses information from

the tandem MS to best assign possible sulfate and acetate modifications. Assignments are made based on confidence values and are used to generate a list of top candidates. This is the first GAG software that requires only the tandem MS for sequence information. While certainly a high-throughput option, the structural assignment conflicts can arise in the form of sulfate loss fragment, internal fragments or random matches. The authors of HS-SEQ note that the software removes the assignments with lower-confidence to resolve conflicting assignments but also believe this may produce false hits when examining samples extracted from biological sources.

1.4 DISSERTATION TOPIC: DEVELOPING SOFTWARE FOR GAG STRUCTURAL CHARACTERIZATION

The fundamental bottleneck that is GAG spectra interpretation still requires a more streamlined and less-specific software suite. Current methodologies are narrow in scope, targeting only specific families with pre-injection derivatization technique or use database-oriented search methods that might be limited to a narrow breath of chain lengths. This dissertation is the story of how we develop a non-specific, non-database driven GAG software suite. Using the MATLAB programming environment, we explore the avenues of *in silico* theoretical fragment prediction and model and use statistical inference methods known as expectation values and survival functions to optimize our structural identification paradigm. Lastly, we apply this method to unknown, heavily sulfated chondroitin sulfates known as decorin GAG and determine their structures.

1.5 REFERENCES

1. Roseman, S., Reflections on glycobiology. *Journal of Biological Chemistry* **2001,** *276* (45), 41527-41542.

2. Pervin, A.; Alhakim, A.; Linhardt, R. J., SEPARATION OF GLYCOSAMINOGLYCAN-DERIVED OLIGOSACCHARIDES BY CAPILLARY ELECTROPHORESIS USING REVERSE POLARITY. *Anal. Biochem.* **1994,** *221* (1), 182-188.

3. Arungundram, S.; Al-Mafraji, K.; Asong, J.; Leach, F. E.; Amster, I. J.; Venot, A.; Turnbull, J. E.; Boons, G. J., Modular Synthesis of Heparan Sulfate Oligosaccharides for Structure-Activity Relationship Studies. *J. Am. Chem. Soc.* **2009,** *131* (47), 17394-17405.

4. Ly, M.; Leach, F. E., III; Laremore, T. N.; Toida, T.; Amster, I. J.; Linhardt, R. J., The proteoglycan bikunin has a defined sequence. *Nature Chemical Biology* **2011,** *7* (11), 827-833.

5. Yu, Y. L.; Duan, J. N.; Leach, F. E.; Toida, T.; Higashi, K.; Zhang, H.; Zhang, F. M.; Amster, I. J.; Linhardt, R. J., Sequencing the Dermatan Sulfate Chain of Decorin. *J. Am. Chem. Soc.* **2017,** *139* (46), 16986-16995.

6. Fraser, J. R. E.; Laurent, T. C.; Laurent, U. B. G., Hyaluronan: Its nature, distribution, functions and turnover. *J Intern Med* **1997,** *242* (1), 27-33.

7. de Sa, V. K.; Rocha, T. P.; Moreira, A. L.; Soares, F. A.; Takagaki, T.; Carvalho, L.; Nicholson, A. G.; Capelozzi, V. L., Hyaluronidases and hyaluronan synthases expression is inversely correlated with malignancy in lung/bronchial pre-neoplastic and neoplastic lesions, affecting prognosis. *Braz J Med Biol Res* **2015,** *48* (11), 1039-1047.

8. Zhang, L. R.; Underhill, C. B.; Chen, L. P., Hyaluronan on the Surface of Tumor-Cells Is Correlated with Metastatic Behavior. *Cancer Res* **1995,** *55* (2), 428-433.

9. Culty, M.; Shizari, M.; Nguyen, H. A.; Clark, R.; Thompson, E. W.; Underhill, C. B., Expression of Cd44 by Human Breast-Cancer Cell-Lines Is Correlated with Hyaluronan Binding, Degradation and Distribution in the Tumor Stroma. *Mol Biol Cell* **1992,** *3*, A320-A320.

10. Mahlbacher, V.; Sewing, A.; Elsasser, H. P.; Kern, H. F., Hyaluronan Is a Secretory Product of Human Pancreatic Adenocarcinoma Cells. *Eur J Cell Biol* **1992,** *58* (1), 28-34.

11. Bourguignon, L. Y. W.; Shiina, M.; Li, J. J., Hyaluronan-CD44 Interaction Promotes Oncogenic Signaling, microRNA Functions, Chemoresistance, and Radiation Resistance in Cancer Stem Cells Leading to Tumor Progression. *Adv Cancer Res* **2014,** *123*, 255-275.

12. Nicholls, M. A.; Fierlinger, A.; Niazi, F.; Bhandari, M., The Disease-Modifying Effects of Hyaluronan in the Osteoarthritic Disease State. *Clin Med Insights-Ar* **2017,** *10*.

13. McCarty, W. J.; Masuda, K.; Sah, R. L., Fluid movement and joint capsule strains due to flexion in rabbit knees. *J Biomech* **2011,** *44* (16), 2761-2767.

14. Moretto, P.; Karousou, E.; Viola, M.; Caon, I.; D'Angelo, M. L.; De Luca, G.; Passi, A.; Vigetti, D., Regulation of Hyaluronan Synthesis in Vascular Diseases and Diabetes. *J Diabetes Res* **2015**.

15. Nelson, R. M.; Venot, A.; Bevilacqua, M. P.; Linhardt, R. J.; Stamenkovic, I., Carbohydrate-protein interactions in vascular biology. *Annu Rev Cell Dev Bi* **1995,** *11*, 601-631.

16. Pomin, V. H., Keratan sulfate: An up-to-date review. *Int J Biol Macromol* **2015,** *72*, 282-289.

17. Linares, P. M.; Chaparro, M.; Algaba, A.; Roman, M.; Arza, I. M.; Santos, F. A.; Ochoa, D.; Guerra, I.; Bermejo, F.; Gisbert, J. P., Effect of Chondroitin Sulphate on Pro-Inflammatory Mediators and Disease Activity in Patients with Inflammatory Bowel Disease. *Digestion* **2015,** *92* (4), 203-210.

18. Singh, J. A.; Noorbaloochi, S.; MacDonald, R.; Maxwell, L. J., Chondroitin for osteoarthritis. *Cochrane Db Syst Rev* **2015,** (1).

19. Kwoh, C. K.; Roemer, F. W.; Hannon, M. J.; Moore, C. E.; Jakicic, J. M.; Guermazi, A.; Green, S. M.; Evans, R. W.; Boudreau, R., Effect of Oral Glucosamine on Joint Structure in Individuals With Chronic Knee Pain. *Arthritis Rheumatol* **2014,** *66* (4), 930-939.

20. Bauerova, K.; Ponist, S.; Kuncirova, V.; Mihalova, D.; Paulovicova, E.; Volpi, N., Chondroitin sulfate effect on induced arthritis in rats. *Osteoarthr Cartilage* **2011,** *19* (11), 1373-1379.

21. de Rezende, M. U.; de Campos, G. C.; Pailo, A. F., Current Concepts in Osteoarthritis. *Acta Ortop Bras* **2013,** *21* (2), 120-122.

22. Durmaz, B., Use of Complementary Medicines for Osteoarthritis. *Turk J Geriatr* **2011,** *14*, 83-88.

23. Fox, B. A.; Stephens, M. M., Glucosamine/Chondroitin/Primorine Combination Therapy for Osteoarthritis. *Drug Today* **2009,** *45* (1), 21-31.

24. Liu, F.; Zhang, N.; Li, Z. J.; Wang, X.; Shi, H. J.; Xue, C. H.; Li, R. W.; Tang, Q. J., Chondroitin sulfate disaccharides modified the structure and function of the murine gut microbiome under healthy and stressed conditions. *Sci Rep-Uk* **2017,** *7*.

25. Mizumoto, S.; Mikami, T.; Yasunaga, D.; Kobayashi, N.; Yamauchi, H.; Miyake, A.; Itoh, N.; Kitagawa, H.; Sugahara, K., Chondroitin 4-O-sulfotransferase-1 is required for somitic muscle development and motor axon guidance in zebrafish. *Biochem. J.* **2009,** *419*, 387-399.

26. Yang, R.; Chen, Y.; Chen, D. Z., Biological functions and role of CCN1/Cyr61 in embryogenesis and tumorigenesis in the female reproductive system. *Mol Med Rep* **2018,** *17* (1), 3-10.

27. Gwon, K.; Kim, E.; Tae, G., Heparin-hyaluronic acid hydrogel in support of cellular activities of 3D encapsulated adipose derived stem cells. *Acta Biomater* **2017,** *49*, 284-295.

28. Mei, L.; Liu, Y. Y.; Zhang, H. J.; Zhang, Z. R.; Gao, H. L.; He, Q., Antitumor and Antimetastasis Activities of Heparin-based Micelle Served As Both Carrier and Drug. *Acs Appl Mater Inter* **2016,** *8* (15), 9577-9589.

29. Chen, Y.; Zhao, J.; Yu, Y. L.; Liu, X. Y.; Lin, L.; Zhang, F. M.; Linhardt, R. J., Antithrombin III-Binding Site Analysis of Low-Molecular-Weight Heparin Fractions. *J Pharm Sci-Us* **2018,** *107* (5), 1290-1295.

30. Beyer, J. T.; Schoeppler, K. E.; Zanotti, G.; Weiss, G. M.; Mueller, S. W.; MacLaren, R.; Fish, D. N.; Kiser, T. H., Antithrombin Administration During Intravenous Heparin Anticoagulation in the Intensive Care Unit: A Single-Center Matched Retrospective Cohort Study. *Clin Appl Thromb-Hem* **2018,** *24* (1), 145-150.

31. Zhao, Y. J.; Singh, A.; Li, L. Y.; Linhardt, R. J.; Xu, Y. M.; Liu, J.; Woods, R. J.; Amster, I. J., Investigating changes in the gas-phase conformation of Antithrombin III upon

binding of Arixtra using traveling wave ion mobility spectrometry (TWIMS). *Analyst* **2015,** *14* (20), 6980-6989.

32. Bandtlow, C. E.; Zimmermann, D. R., Proteoglycans in the developing brain: New conceptual insights for old proteins. *Physiol Rev* **2000,** *80* (4), 1267-1290.

33. Zhao, Y. J.; Singh, A.; Xu, Y. M.; Zong, C. L.; Zhang, F. M.; Boons, G. J.; Liu, J.; Linhardt, R. J.; Woods, R. J.; Amster, I. J., Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (1), 96-109.

34. Harvey, D. J., Derivatization of carbohydrates for analysis by chromatography; electrophoresis and mass spectrometry. *J Chromatogr B* **2011,** *879* (17-18), 1196-1225.

35. Kuster, B.; Hunter, A. P.; Wheeler, S. F.; Dwek, R. A.; Harvey, D. J., Structural determination of N-linked carbohydrates by matrix-assisted laser desorption/ionization-mass spectrometry following enzymatic release within sodium dodecyl sulphate-polyacrylamide electrophoresis gels: application to species-specific glycosylation of alpha1-acid glycoprotein. *Electrophoresis* **1998,** *19* (11), 1950-9.

36. Zhou, W.; Hakansson, K., Structural Characterization of Carbohydrates by Fourier Transform Tandem Mass Spectrometry. *Curr Proteomics* **2011,** *8* (4), 297-308.

37. Zaia, J., Mass spectrometry of oligosaccharides. *Mass Spectrom Rev* **2004,** *23* (3), 161-227.

38. Zaia, J., Mass spectrometry and glycomics. *OMICS* **2010,** *14* (4), 401-18.

39. Harvey, D. J., Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics* **2001,** *1* (2), 311-28.

40. Wuhrer, M., Glycomics using mass spectrometry. *Glycoconj J* **2013,** *30* (1), 11-22.

41. Park, Y.; Lebrilla, C. B., Application of Fourier transform ion cyclotron resonance mass spectrometry to oligosaccharides. *Mass Spectrom Rev* **2005,** *24* (2), 232-64.

42. Harvey, D. J., Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrom Rev* **1999,** *18* (6), 349-450.

43. Karas, M.; Bahr, U.; Dulcks, T., Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresenius J Anal Chem* **2000,** *366* (6-7), 669-76.

44. Shi, X.; Shao, C.; Mao, Y.; Huang, Y.; Wu, Z. L.; Zaia, J., LC-MS and LC-MS/MS studies of incorporation of 34SO3 into glycosaminoglycan chains by sulfotransferases. *Glycobiology* **2013,** *23* (8), 969-79.

45. Auray-Blais, C.; Lavoie, P.; Zhang, H.; Gagnon, R.; Clarke, J. T.; Maranda, B.; Young, S. P.; An, Y.; Millington, D. S., An improved method for glycosaminoglycan analysis by LC-MS/MS of urine samples collected on filter paper. *Clin Chim Acta* **2012,** *413* (7-8), 771-8.

46. Staples, G. O.; Bowman, M. J.; Costello, C. E.; Hitchcock, A. M.; Lau, J. M.; Leymarie, N.; Miller, C.; Naimy, H.; Shi, X.; Zaia, J., A chip-based amide-HILIC LC/MS platform for glycosaminoglycan glycomics profiling. *Proteomics* **2009,** *9* (3), 686-95.

47. Domon, B.; Costello, C. E., A SYSTEMATIC NOMENCLATURE FOR CARBOHYDRATE FRAGMENTATIONS IN FAB-MS MS SPECTRA OF GLYCOCONJUGATES. *Glycoconjugate J.* **1988,** *5* (4), 397-409.

48. Galeotti, F.; Volpi, N., Online Reverse Phase-High-Performance Liquid Chromatography-Fluorescence Detection-Electrospray Ionization-Mass Spectrometry Separation and Characterization of Heparan Sulfate, Heparin, and Low-Molecular Weight-Heparin Disaccharides Derivatized with 2-Aminoacridone. *Analytical Chemistry* **2011,** *83* (17), 6770-6777.

49. Huang, R. R.; Pomin, V. H.; Sharp, J. S., LC-MS (n) Analysis of Isomeric Chondroitin Sulfate Oligosaccharides Using a Chemical Derivatization Strategy. *Journal of the American Society for Mass Spectrometry* **2011,** *22* (9), 1577-1587.

50. Leach, F. E.; Ly, M.; Laremore, T. N.; Wolff, J. J.; Perlow, J.; Linhardt, R. J.; Amster, I. J., Hexuronic Acid Stereochemistry Determination in Chondroitin Sulfate Glycosaminoglycan Oligosaccharides by Electron Detachment Dissociation. *Journal of the American Society for Mass Spectrometry* **2012,** *23* (9), 1488-1497.

51. Agyekum, I.; Patel, A. B.; Zong, C. L.; Boons, G. J.; Amster, I. J., Assignment of hexuronic acid stereochemistry in synthetic heparan sulfate tetrasaccharides with 2-O-sulfo

uronic acids using electron detachment dissociation. *Int. J. Mass Spectrom.* **2015,** *390*, 163-169.

52. Zaia, J., Principles of Mass Spectrometry of Glycosaminoglycans. *Journal of Biomacromolecular Mass Spectrometry* **2005,** *1* (1), 3-36.

53. Kailemia, M. J.; Li, L. Y.; Ly, M.; Linhardt, R. J.; Amster, I. J., Complete Mass Spectral Characterization of a Synthetic Ultralow-Molecular-Weight Heparin Using Collision-Induced Dissociation. *Analytical Chemistry* **2012,** *84* (13), 5475-5478.

54. Cody, R. B.; Burnier, R. C.; Freiser, B. S., COLLISION-INDUCED DISSOCIATION WITH FOURIER-TRANSFORM MASS-SPECTROMETRY. *Analytical Chemistry* **1982,** *54* (1), 96-101.

55. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007,** *4* (9), 709-712.

56. Kailemia, M. J.; Patel, A. B.; Johnson, D. T.; Li, L. Y.; Linhardt, R. J.; Amster, I. J., Differentiating chondroitin sulfate glycosaminoglycans using collision-induced dissociation; uronic acid cross-ring diagnostic fragments in a single stage of tandem mass spectrometry. *European Journal of Mass Spectrometry* **2015,** *21* (3), 275-285.

57. Kruger, N. A.; Zubarev, R. A.; Horn, D. M.; McLafferty, F. W., Electron capture dissociation of multiply charged peptide cations. *International Journal of Mass Spectrometry* **1999,** *185*, 787-793.

58. Zubarev, R. A.; Kruger, N. A.; Fridriksson, E. K.; Lewis, M. A.; Horn, D. M.; Carpenter, B. K.; McLafferty, F. W., Electron capture dissociation of gaseous multiply-charged proteins is favored at disulfide bonds and other sites of high hydrogen atom affinity. *J. Am. Chem. Soc.* **1999,** *121* (12), 2857-2862.

59. Budnik, B. A.; Haselmann, K. F.; Zubarev, R. A., Electron detachment dissociation of peptide di-anions: an electron-hole recombination phenomenon. *Chemical Physics Letters* **2001,** *342* (3-4), 299-302.

60. Cooper, H. J.; Hakansson, K.; Marshall, A. G., The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom Rev* **2005,** *24* (2), 201-222.

61. Agyekum, I.; Zong, C. L.; Boons, G. J.; Amster, I. J., Single Stage Tandem Mass Spectrometry Assignment of the C-5 Uronic Acid Stereochemistry in Heparan Sulfate Tetrasaccharides using Electron Detachment Dissociation. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (9), 1741-1750.

62. Huang, Y.; Yu, X.; Mao, Y.; Costello, C. E.; Zaia, J.; Lin, C., De Novo Sequencing of Heparan Sulfate Oligosaccharides by Electron-Activated Dissociation. *Analytical Chemistry* **2013,** *85* (24), 11979-11986.

63. Leach, F. E.; Riley, N. M.; Westphall, M. S.; Coon, J. J.; Amster, I. J., Negative Electron Transfer Dissociation Sequencing of Increasingly Sulfated Glycosaminoglycan Oligosaccharides on an Orbitrap Mass Spectrometer. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (9), 1844-1854.

64. Wolff, J. J.; Leach, F. E.; Laremore, T. N.; Kaplan, D. A.; Easterling, M. L.; Linhardt, R. J.; Amster, I. J., Negative Electron Transfer Dissociation of Glycosaminoglycans. *Analytical Chemistry* **2010,** *82* (9), 3460-3466.

65. Cannon, J. R.; Carnmarata, M. B.; Robotham, S. A.; Cotham, V. C.; Shaw, J. B.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., Ultraviolet Photodissociation for Characterization of Whole Proteins on a Chromatographic Time Scale. *Analytical Chemistry* **2014,** *86* (4), 2185-2192.

66. Shaw, J. B.; Li, W. Z.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. *J. Am. Chem. Soc.* **2013,** *135* (34), 12646-12651.

67. Devakumar, A.; Thompson, M. S.; Reilly, J. P., dFragmentation of oligosaccharide ions with 157 nm vacuum ultraviolet light. *Rapid Communications in Mass Spectrometry* **2005,** *19* (16), 2313-2320.

68. Ko, B. J.; Brodbelt, J. S., 193 nm Ultraviolet Photodissociation of Deprotonated Sialylated Oligosaccharides. *Analytical Chemistry* **2011,** *83* (21), 8192-8200.

69. Ruhaak, L. R.; Zauner, G.; Huhn, C.; Bruggink, C.; Deelder, A. M.; Wuhrer, M., Glycan labeling strategies and their use in identification and quantification. *Analytical and Bioanalytical Chemistry* **2010,** *397* (8), 3457-3481.

70. Yang, B.; Chang, Y. D.; Weyers, A. M.; Sterner, E.; Linhardt, R. J., Disaccharide analysis of glycosaminoglycan mixtures by ultra-high-performance liquid chromatography-mass spectrometry. *J Chromatogr A* **2012,** *1225*, 91-98.

71. Langeslay, D. J.; Urso, E.; Gardini, C.; Naggi, A.; Torri, G.; Larive, C. K., Reversed-phase ion-pair ultra-high-performance-liquid chromatography-mass spectrometry for fingerprinting low-molecular-weight heparins. *J Chromatogr A* **2013,** *1292*, 201-210.

72. Du, J. Y.; Chen, L. R.; Liu, S.; Lin, J. H.; Liang, Q. T.; Lyon, M.; Wei, Z., Ion-pairing liquid chromatography with on-line electrospray ion trap mass spectrometry for the structural analysis of N-unsubstituted heparin/heparan sulfate. *J Chromatogr B* **2016,** *1028*, 71-76.

73. Yang, B.; Weyers, A.; Baik, J. Y.; Sterner, E.; Sharfstein, S.; Mousa, S. A.; Zhang, F. M.; Dordick, J. S.; Linhardt, R. J., Ultra-performance ion-pairing liquid chromatography with on-line electrospray ion trap mass spectrometry for heparin disaccharide analysis. *Anal. Biochem.* **2011,** *415* (1), 59-66.

74. Staples, G. O.; Bowman, M. J.; Costello, C. E.; Hitchcock, A. M.; Lau, J. M.; Leymarie, N.; Miller, C.; Naimy, H.; Shi, X. F.; Zaia, J., A chip-based amide-HILIC LC/MS platform for glycosaminoglycan glycomics profiling. *Proteomics* **2009,** *9* (3), 686-695.

75. Huang, Y.; Shi, X. F.; Yu, X.; Leymarie, N.; Staples, G. O.; Yin, H. F.; Killeen, K.; Zaia, J., Improved Liquid Chromatography-MS/MS of Heparan Sulfate Oligosaccharides via Chip-Based Pulsed Makeup Flow. *Analytical Chemistry* **2011,** *83* (21), 8222-8229.

76. Karlsson, N. G.; Schulz, B. L.; Packer, N. H.; Whitelock, J. M., Use of graphitised carbon negative ion LC-MS to analyse enzymatically digested glycosaminoglycans. *J Chromatogr B* **2005,** *824* (1-2), 139-147.

77. Gill, V. L.; Wang, Q.; Shi, X. F.; Zaia, J., Mass Spectrometric Method for Determining the Uronic Acid Epimerization in Heparan Sulfate Disaccharides Generated Using Nitrous Acid. *Analytical Chemistry* **2012,** *84* (17), 7539-7546.

78. Lin, L.; Liu, X. Y.; Zhang, F. M.; Chi, L. L.; Amster, I. J.; Leach, F. E.; Xia, Q. W.; Linhardt, R. J., Analysis of heparin oligosaccharides by capillary electrophoresis-negative-

ion electrospray ionization mass spectrometry. *Analytical and Bioanalytical Chemistry* **2017,** *409* (2), 411-420.

79. Sanderson, P.; Stickney, M.; Leach, F. E.; Xia, Q. W.; Yu, Y. L.; Zhang, F. M.; Linhardt, R.; Amster, I. J., Heparin/heparan sulfate analysis by covalently modified reverse polarity capillary zone electrophoresis-mass spectrometry. *J Chromatogr A* **2018,** *1545*, 75-83.

80. Singh, A.; Kett, W. C.; Severin, I. C.; Agyekum, I.; Duan, J. N.; Amster, I. J.; Proudfoot, A. E. I.; Coombe, D. R.; Woods, R. J., The Interaction of Heparin Tetrasaccharides with Chemokine CCL5 Is Modulated by Sulfation Pattern and pH. *Journal of Biological Chemistry* **2015,** *290* (25), 15421-15436.

81. Kailemia, M. J.; Park, M.; Kaplan, D. A.; Venot, A.; Boons, G. J.; Li, L. Y.; Linhardt, R. J.; Amster, I. J., High-Field Asymmetric-Waveform Ion Mobility Spectrometry and Electron Detachment Dissociation of Isobaric Mixtures of Glycosaminoglycans. *Journal of the American Society for Mass Spectrometry* **2014,** *25* (2), 258-268.

82. Seo, Y.; Andaya, A.; Leary, J. A., Preparation, Separation, and Conformational Analysis of Differentially Sulfated Heparin Octasaccharide Isomers Using Ion Mobility Mass Spectrometry. *Analytical Chemistry* **2012,** *84* (5), 2416-2423.

83. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M., The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biological Chemistry* **2012,** *393* (11), 1357-1362.

84. Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **2003,** *17* (20), 2337-2342.

85. Clauser, K. R.; Baker, P.; Burlingame, A. L., Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS MS and database searching. *Analytical Chemistry* **1999,** *71* (14), 2871-2882.

86. Saad, O. M.; Leary, J. A., Heparin sequencing using enzymatic digestion and ESI-MSn with HOST: A heparin/HS oligosaccharide sequencing tool. *Analytical Chemistry* **2005,** *77* (18), 5902-5911.

87. Chiu, Y. L.; Huang, R. R.; Orlando, R.; Sharp, J. S., GAG-ID: Heparan Sulfate (HS) and Heparin Glycosaminoglycan High-Throughput Identification Software. *Mol. Cell. Proteomics* **2015,** *14* (6), 1720-1730.

88. Hu, H.; Huang, Y.; Mao, Y.; Yu, X.; Xu, Y. M.; Liu, J.; Zong, C. L.; Boons, G. J.; Lin, C.; Xia, Y.; Zaia, J., A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra. *Mol. Cell. Proteomics* **2014,** *13* (9), 2490-2502.

CHAPTER 2

AN AUTOMATED, HIGH-THROUGHPUT METHOD FOR INTERPRETING THE TANDEM MASS SPECTRA OF GLYCOSAMINOGLYCANS*

Duan, J.N.; Amster, I. J. *J. Amer. Soc. Mass. Spec.* **2018.** *In press.*

*Reprinted with permission from Journal of the American Society of Mass Spectrometry.

2.1 ABSTRACT

The biological interactions between glycosaminoglycans (GAGs) and other biomolecules are heavily influenced by structural features of the glycan. The structure of GAGs can be assigned using tandem mass spectrometry (MS^2), but analysis of these data, to date, requires manually interpretation, a slow process that presents a bottleneck to the broader deployment of this approach to solving biologically relevant problems. Automated interpretation remains a challenge, as GAG biosynthesis is not template-driven, and therefore one cannot predict structures from genomic data, as is done with proteins. The lack of a structure database, a consequence of the non-template biosynthesis, requires a *de novo* approach to interpretation of the mass spectral data. We propose a model for rapid, high-throughput GAG analysis by using an approach in which candidate structures are scored for the likelihood that they would produce the features observed in the mass spectrum. To make this approach tractable, a genetic algorithm is used to greatly reduce the search-space of isomeric structures that are considered. The amount of time required for analysis is significantly reduced compared to an approach in which every possible isomer is considered and scored. The model is coded in a software package using the MATLAB environment. This approach was tested on tandem mass spectrometry data for long chain, moderately sulfated chondroitin sulfate oligomers that were derived from the proteoglycan bikunin. The bikunin data was previously interpreted manually. Our approach examines glycosidic fragments to localize SO_3 modifications to specific residues and yields the same structures reported in literature, only much more quickly.

2.2 INTRODUCTION

Glycosaminoglycans (GAGs) are linear, polydisperse carbohydrates consisting of a repeating uronic sugar and amino sugar copolymer. GAGs serve a multitude of roles in biology including cell-cell and cell-matrix interactions, generation of energy, changes in proteins binding conformation, and molecular recognition[1-3]. Certain GAGs have also been observed as potential biomarkers for disease states[4]. The degree of GAG-protein binding has been shown to be highly dependent on their structure and, more specifically, the position of modifications within their generic repeating copolymer chain [5-6].

Despite the simple polymeric backbone in GAGs, a single sugar residue can exhibit varying levels of three key modifications, namely O-sulfation, N-deacetylation/sulfation, and uronic sugar stereochemistry[2]. Moreover, the biosynthesis of GAGs is not template driven, resulting in non-uniform dispersion of these modifications across the chain[7-8]. Database-derived approaches are widely used for protein mass spectra assignment (either top-down or bottom-up) due to the predictability of amino acid sequences from genome sequences, but fail when applied to biomolecules whose production is not template-derived [9-10]. In contrast to the approaches that are successful for protein/peptide analysis, a *de novo* approach is required for the computer-based analysis of the tandem mass spectra of GAGs.

Considerable progress has been made in GAG analysis using mass spectrometry [1, 11]. At the MS^1 level, a parts-per-million accurate mass measurement, using high resolution instruments such as Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS), allows assignment of composition, from which GAG chain length, number of modifications and types of modification can be assigned [12]. Tandem MS (MS^2) of GAGs using various ion activation methods, such as collision-induced dissociation (CID)[13-15],

infrared multiphoton dissociation [16-19], electron-detachment dissociation (EDD)[16, 18-24], and negative-electron transfer dissociation (NETD) [25-27] yields structurally informative fragment ions [28]. Glycosidic bond fragmentation provides monosaccharide composition, while cross-ring fragmentation is used to assign the location of modifications within each residue [29]. Because this is a *de novo* analytical approach, complete structure analysis requires an information-rich mass spectrum that contains sufficient fragment peaks to fully assign all the variable features. Recent developments in ion activation for GAGs has led to a variety of approaches to produce informative MS^2 spectra [21, 23, 28, 30]. However, the interpretation of the such complex mass spectra is generally a tedious manual process that relies upon the expertise of the data analyst. A better understanding of the structural features that promote GAG activity would benefit from an automated, accurate and high-throughput analytical process.

The complexity of the data sets and the time required for analysis increases dramatically as the chain length and the number of modifications increase. Two families of GAGs, heparin/heparan sulfate (Hp/HS) and chondroitin/dermatan sulfate (CS/DS), often contain large numbers of labile sulfate modifications. For these compounds, conventional MS^2 methods are often inadequate for complete structural determination, either because they do not produce a comprehensive set of fragment ions required to assign all variable features, or because they lead to decomposition products that confound the analysis [8, 31]. For example, fragmentation can be accompanied by decomposition of sulfo modifications, producing peaks that are reduced in mass by multiples of 80 mass units, but match the mass of standard glycosidic fragments of their counterparts with fewer sulfate modifications [28, 32]. If one does not recognize the peaks that arise from such decomposition,

incorrect structural assignments will result. Common *de novo* strategies that have been successful for protein sequencing [25, 33-35] will inevitably be exposed to substantially more false positives due to the high-likelihood of SO_3 loss fragments in GAG MS and MS^2. Na^+/H^+ exchange has been shown to decrease SO_3 loss and makes characterization of highly sulfate species possible [30], however SO_3 loss is almost always observed in MS^2 spectra.

An alternative to the above approach to interpretation is to generate a list of possible fragment peaks for a candidate structure, and to score the match with the experimental data. This process can be repeated for all possible isomers having a given elemental composition. Comparison of the experimental MS^2 against the theoretical fragment list allows us to rank each permutation based on closeness-of-fit to the experimental results. This method becomes impractical to perform manually when the number of possible permutations for a composition exceeds the capability to examine the data. For example, Arixtra, a heparin with 5 monosaccharides, is the largest highly sulfated GAG to have complete mass spectral characterization [30]. The number of total possible permutations for a GAG scale logarithmically with the respect to chain length. For both chondroitin/dermatan sulfate and heparan sulfate/heparin, the number of permutations based on chain length and number of modifications is calculated as *n*-choose-*k* combinations, where *n* is the number of possible modifiable sites and *k* is the number of modifications:

$$(eq.1) \quad N_{total} \propto \log N_{chain\ lengt}$$

$$(eq.2) \quad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Tools for comparison of user-input structures with fragment peaks from tandem MS have been developed[12, 36-37], but the requirement for a known starting structure limit applicability for high-throughput analysis.

To address this bottleneck for high-throughput sequencing of GAGs, efforts in computer-assisted methods look to improve upon the speed of analysis and to reduce the amount of user-input and supervision. Several software packages have been developed to overcome modern challenges in GAG analysis although a few require addition steps at the experimental level for optimal software performance. HOST [38] is a computational tool designed for sequencing heparin/HS oligosaccharides using enzymatic digestion combined with ESI-MS^n. The method scores and returns the best matching sequences of GAGs based on disaccharide composition analysis, yielding predicted compositions and calculating expected fragmentation patterns *in silico*. Comparisons of theoretical fragments can then be compared to fragmentation of heparin/HS oligosaccharide MS^n data and is scored to return the most likely sequence. However, disaccharide analysis requires complete enzymatic digestion of the GAG using heparin lyase I, II and III over multiple hours of incubation (16 h), limiting the method's overall speed and applicability in a high-throughput GAG analysis platform.

Another piece of software known as GAG-ID [39] has been shown to discriminate and identify 21 synthetic tetrasaccharides eluted from LC-MS/MS using a scoring system based on peak intensities. It is the first of its kind to automated the interpretation of mixtures when coupled to LC-MS/MS but require complete chemical derivatization of the GAG by replacing all labile sulfate modifications with more stable acetyl groups. Much like HOST, derivatization may not be a viable option for universal GAG analysis.

HS-SEQ [40] is a *de novo* GAG sequencing computation framework that has been used to automate the structural identification of HS of dp4, 5, 6, 8 and 15. The method determines a precursor sequence (unmodified GAG backbone) and uses information from the tandem MS to best assign possible sulfate and acetate modifications. Assignments are made based on confidence values and are used to generate a list of top candidates. This is the first GAG software that requires only the tandem MS for sequence information. While certainly a high-throughput option, the structural assignment conflicts can arise in the form of sulfate loss fragment, internal fragments or random matches. The authors of HS-SEQ note that the software removes the assignments with lower-confidence to resolve conflicting assignments but also believe this may produce false hits when examining samples extracted from biological sources.

The software developed in our laboratory is designed to sequence GAGs of indefinite length by comparing fragments of theoretical structures (*in silico*) against experimental data without the need for construction of a database, instead using a genetic algorithm optimization technique to limit the number of permutations while keeping analysis time to a maximum of a few minutes. The method assigns structures based on greatest likelihood using fragment ion products as a critical parameter for the genetic algorithm fitness criterion. Fragments that are in direct conflict with the highest scoring structure(s) are not discarded but reviewed again for possible additional components. We have tested this approach on MS^2 data from intact CS chains released from the proteoglycan, bikunin. These chains vary in length from 27-43 saccharide residues, and vary in the degree of O-sulfo modification from 4 to 7, and thus represent a challenging test of this automated procedure.

2.3 EXPERIMENTAL METHODS

Mass spectrometry analysis. Bikunin GAG MS and MS^2 data reported in [41] was used as a proof-of-principle data set for the purposes of testing genetic algorithm efficacy. The monoisotopic peaks were selected via the SNAP algorithm from Bruker DataAnalysis software. Analysis of the MS^2 was performed with the software alone and with no user supervision or assistance.

Computational methods. MS^1 analysis of parent ion mass is performed using a composition assignment software module written in the MATLAB coding environment. Monoisotopic peaks and charge states are acquired from Bruker DataAnalysis and deconvoluted to a neutral mass. A composition is derived from one or more neutral mass(es) by searching a data matrix of possible chain lengths, degrees of sulfation, deacetylation, and sodium/hydrogen exchange. The user input also includes the possibility of reducing end modifications, and nonreducing ends that can terminate in unsaturated uronic acids, as is common in enzymatically produced GAG oligomers. Theoretical neutral masses in the spreadsheet are compared against user specified masses with a user-defined mass tolerance. The sequences that match are then used for performing the MS^2 analysis.

For MS^2 assignment, we implement a genetic algorithm based on fundamental aspects common to all genetic algorithms[42-44]. For MS^2 analysis, the software uses a binary vector to represent glycan structures where on-bits denote an occupied site of SO_3 modification. The first step generates two glycan structures at random that fit the expected composition (*initialization step*) and then proceeds to "breed" these structures into a new generation of candidates (*crossover step*). The new generation also is subject to potential

mutations in their structure in the form of exchanges between their on and off-bits (*mutation step*) in an effort to avoid converging upon a local maximum. Theoretical structures created in the crossover and mutation steps are then tested against the experimental MS^2 data where the score of each structure is determined based on a closeness-of-fit paradigm (*fitness*). The scoring system is subject to various factors that will be discussed in detail in future papers. In the case of bikunin, the score of a structure is a naïve model that determines the top candidate based on the number of matching glyocosidic fragments. The primary three steps (crossover, mutation and fitness) are iterated until the maximum fitness value does not change after numerous cycles. The number of iterations required before termination of the algorithm can be defined by the user but is defaulted at a value of 3. The structure(s) containing the highest scores are then examined using additional data interpretation tools that assign fragment peak masses alongside their charge, intensity and mass error (in ppm).

Experimental MS^2 data collected by FT-ICR is extracted from Bruker Apex user interface software using the SNAP peak-picking algorithm. Monoisotopic peak masses and intensities are extracted in the form of comma-separate value (.csv) files. MATLAB software prompts the user for a .csv file containing mass-to-charge in column 1 and intensity in column 2, with mass-to-charge sorted in ascending order. Parent ion mass and charge must be provided by the user as well as mass information pertaining to a linker region mass on the reducing end. Composition details (chain length and numbers of: sulfation, n-acetylation, Na-H exchange) are calculated from a composition calculation module and then given to the software in the preliminary step before initializing the genetic algorithm.

For bikunin proteoglycan a linker mass of 641.1473 (Gal4S-Gal-Xyl-Serine) was used with the remainder of the bikunin chain length represented as a binary vector. Software integrates separate functional modules to perform mass calculations of theoretical fragment ions, performing standard genetic algorithm features, and scoring theoretical structures against experimental data.

2.4 RESULTS AND DISCUSSION

As GAG chain length and modification increases, the number of possible structural permutations exceeds a value suitable for practical, computationally efficient search methods. For the chondroitin sulfate oligomers studied here, the number of structural possibilities is as large as 3.7E22 for an oligomer of length 50 (eq. 2). The number of possibilities is narrowed down when composition can be assigned and the number of known sulfate modifications is determined. While the paradigm for comparing theoretical structures against experimental data can differ, a minimum number of elements such as fragment type, fragment intensity and sequence coverage must be considered for complete GAG characterization [45]. Thus, instead of trying to shortcut these facets of analysis, we chose an approach that reduces the total search space. Hundreds of millions of structures may exist for a specific GAG composition but for a pure sample only one of these structures is a valid assignment. The impracticality of searching through a massive number of incorrect structures is reduced dramatically when a genetic algorithm search heuristic is applied [44].

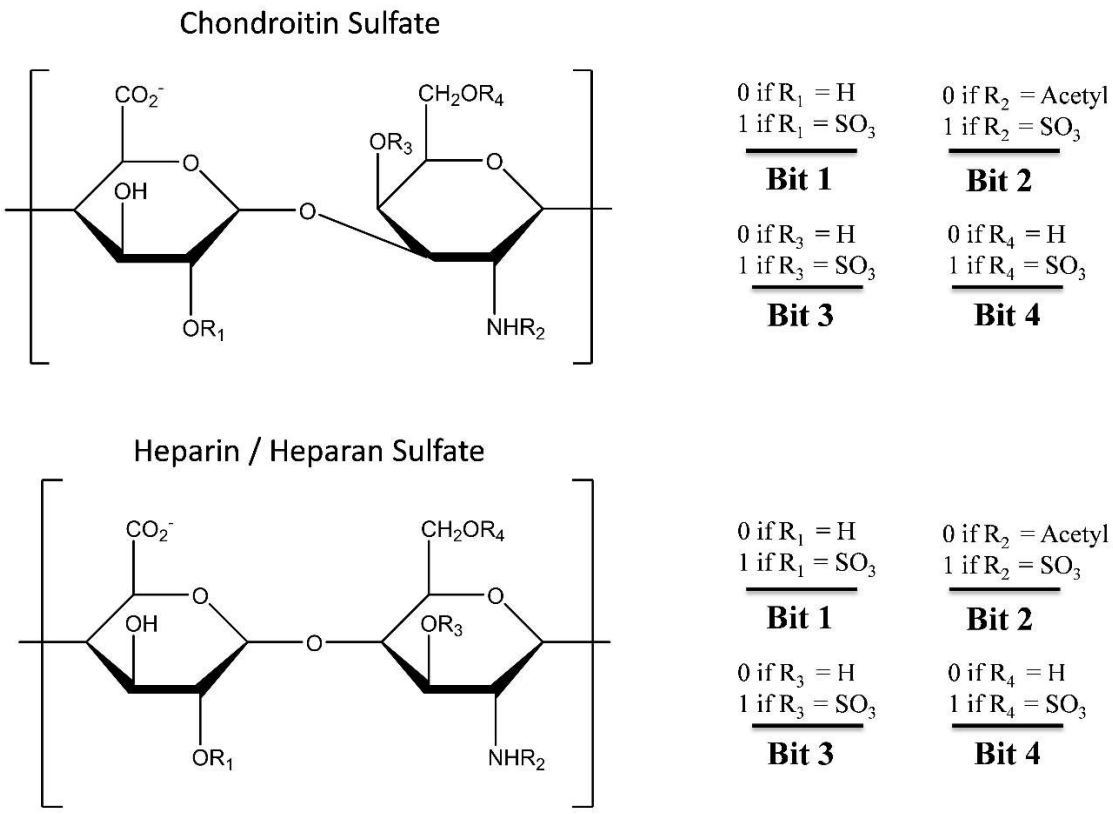
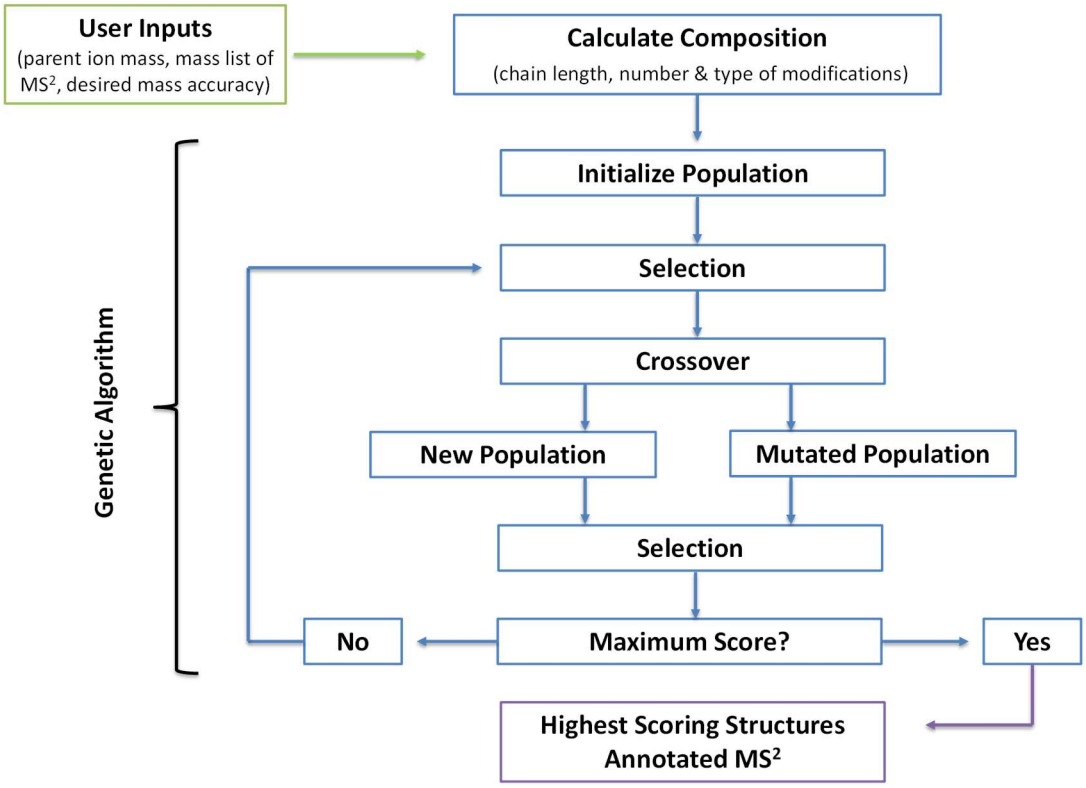


Figure 2.1 4-bit binary representation for both CS and HS/Hp glycan disaccharides. Each bit is turned on (assigned 1) if a modification is present and off (assigned 0) if the R-group is a hydrogen. Bit 2 represents R_2 which has an acetyl modification instead of a hydrogen for an off-bit assignment. In the case of HS where the free-amine is possible, a different numeral can be used to represent the absence of SO_3 and acetylation. Additional bits can be introduced so serve as negative control bits as well as a representation for the uronic sugar stereochemistry.

The genetic algorithm is an optimization tool that has been used for a wide variety of applications[46-51]. It mimics the evolutionary process, by using a survival of the fittest mechanism that quickly eliminates large groups of candidates from a pool if they share a feature that does not meet a specific set of criteria [44]. Here we examine the application of this approach to GAG MS^2 analysis. We have developed software in the MATLAB coding

environment that utilizes the genetic algorithm. GAG sequences are expressed as a binary code where on-bits (1's) and off-bits (0's) represent the presence or absence of modifications, respectively and can be applied to both CS/DS and HS/Hp GAG classes, Figure 2.1 [42-43]. The binary sequence is shortened or lengthened to accommodate the appropriate composition calculated from the parent-ion mass. The number of on and off bits in the genome is also adjusted based on the number of modifications observed. The final structure is determined via a genetic algorithm, the workflow for which is shown in Figure 2.2.



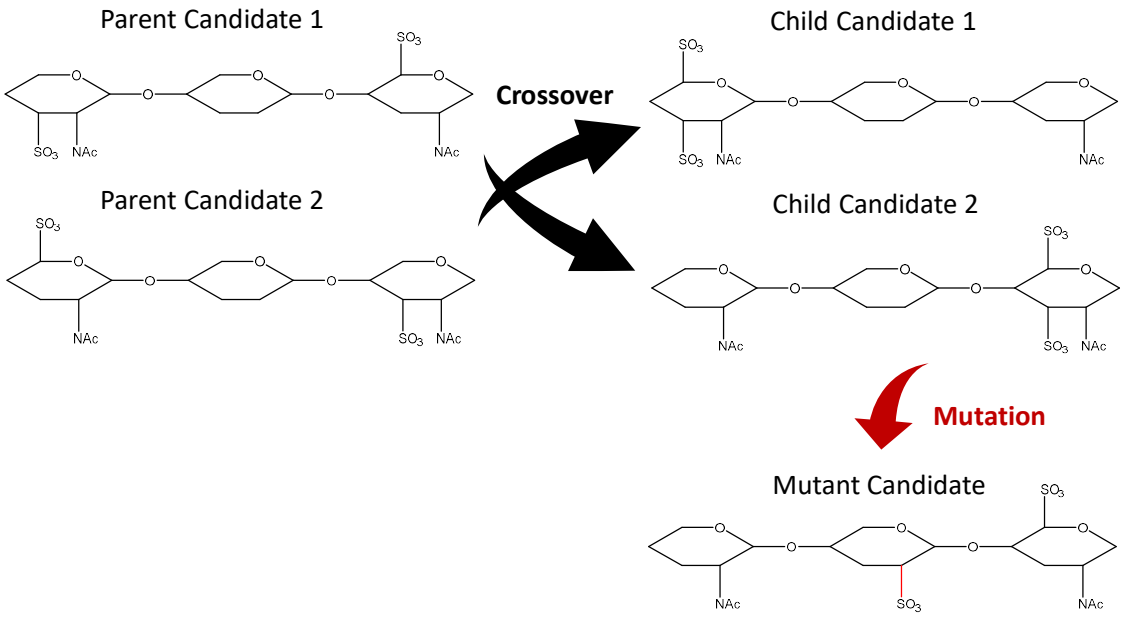


Figure 2.2a. Workflow for our MATLAB software. User is asked to input three pieces of information for the software: parent ion mass, mass list from MS^2 (charge state deconvolution will be automated), and desired mass accuracy for composition assignment and fragment matching (in ppm). The software automates the remaining steps and calculates compositions from the parent ion mass and generates a list of optimized structures using a genetic algorithm. (User provided information is highlighted in the green box. Automated features are highlighted in blue. Software output is shown in purple.) **2.2b.** A demonstration of how genetic operators work on glycan structures. Child candidate modification positions are limited to the modification position of their parents. Mutations, however, and not dependent on parent candidate structure.

Improvements in analysis time and search space reduction can be observed using CID MS^2 data from several fractions of intact CS chains for the proteoglycan bikunin [41]. The advantage of using these data is threefold. First, the mass spectra are rich in structurally informative fragments. Structural assignment of bikunin from MS^2 was done previously with manual *de novo* analysis of these fragments. Software suitable for analysis

should make the same assignments using these fragments without any user supervision. A second advantage is that modifications are limited to a single sulfate group per disaccharide. Sulfate modifications have been shown to only occur on the 4-O position of the amino sugar using enzymatic disaccharide analysis. Reducing the total number of possible modification diminishes the search space dramatically. For example, a CS dp43 with 5 sulfate groups has 20,349 possible structures when only examining the occupancy of the 4-O position but 5,949,147 possible structures when every sulfate position (2-O, 4-O, 6-O) is taken into consideration. A simplified search space allows us to demonstrate proof of principle while still maintaining computational efficiency. Finally, the structures of bikunin fractions have been manually verified and reported in the literature [41]. A common motif among bikunin fractions was observed after manual sequence analysis. We were particularly interested to see if the unsupervised approach with our software also yielded these same patterns. Candidate structures of bikunin GAGs produced in the genetic algorithm cycles are assigned scores based on the number of matched glycosidic fragments in the experimental data. The fitness of a candidate structure is determined using three separate tiers of scoring:

$$(eq.3) \quad f_1 = \sum_{i=1}^{dp} N_{RE} - \sum_{i=1}^{dp} N_{RE+SO}$$

$$(eq.4) \quad f_2 = \sum_{i=1}^{dp} N_{NRE} - \sum_{i=1}^{dp} N_{NRE+S}$$

$$(eq.5) \quad f_3 = \sum_{i=1}^{dp} I_{glyc}$$

Unambiguous mass tags such as the linker region dictate that greater emphasis should be placed on the reducing end (Y and Z fragments) and provide a more valid structural assignment. The primary fitness of a score is therefore based on its calculated f_1 value, which considers the number of glycosidic fragments from the reducing end (N_{RE}) that are matched in the experimental data. The software then checks to see if any match is potentially a sulfate decomposition peak by adding the mass of an SO_3-H exchange (79.9568 Da) and searches the experimental data again for a matching mass. The value of f_1 is then reduced by the number of peaks determined to be a product of sulfate decomposition (N_{RE+SO3}).

If the value of f_1 is tied among multiple structures, a secondary ranking is then determined with f_2, the value of which is based on the number of glycosidic matches from the non-reducing end (B and C fragments). In similar fashion to calculating f_1, considerations for potential sulfate decompositions are considered. Non-reducing end fragments are a tier below reducing end fragments since they could potentially match internal fragments due to the lack of an unambiguous mass tag. Incorrect assignment of internal fragments as non-reducing end fragments limits the validity of assignment.

A tertiary score f_3 is used after matching glycosidic fragments from both reducing and non-reducing ends. Typically, a small selection of candidate structures (2-4) may end up with equal f_1 and f_2 values, in which case the summation of the intensities of all matched glycosidic fragments is the tiebreaker. This simple algorithm can and should be continuously fine-tuned for other purposes as software development continues but is sufficient for proof-of-principle purposes.

11 bikunin samples of different compositions were tested using the genetic algorithm. Of these 11, the single highest scoring candidate of the genetic algorithm for 9 of these samples matched the structures reported in literature. Without user supervision, the genetic algorithm results also reaffirm the common bikunin motif reported in literature [41], figure 2.3. For the remaining 2 samples, the genetic algorithm software reported multiple top-scoring candidates. MS^2 data for these two samples could not unambiguously differentiate these structures; however, the structures reported in literature for these samples were present among the top-scoring candidates. This highlights the importance of data quality for optimal software performance. A lack of informative fragmentation peaks can result in structural ambiguities, but information-rich mass spectra can be interpreted with minimal trouble. However, a genetic algorithm approach has no theoretical minimum for data quality. Spectra not containing sufficient fragmentation for complete glycan characterization can still be interpreted based on available fragment ions and a partial sequence can be generated. Although the spectral quality of bikunin GAG tandem MS are high, more complex and longer chain intact GAGs of proteoglycans may yield less than the full suite of fragments necessary for complete sequencing. In this event, our approach can still be used to determine some portion of the overall glycan structure, as has been done recently for decorin glycans [52].

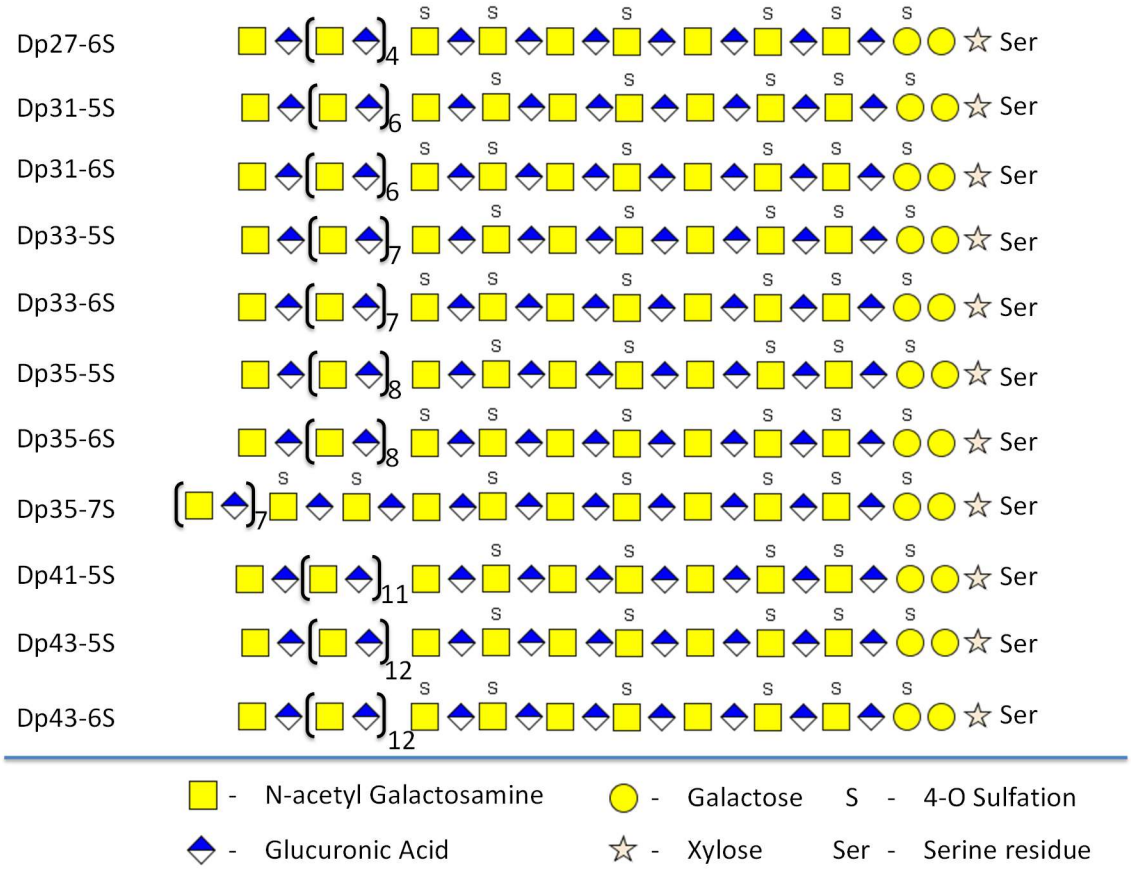


Figure 2.3. A list of the highest scoring structures for all MS^2 collected on FT-ICR using the genetic algorithm. The structures provided by the genetic algorithm match ones reported in literature. The conserved sulfation pattern of bikunin is also observed. For structures dp43-5S and dp43-6S, three structures are tied for highest scores. Alternate structures for these chain lengths are shown in figure.

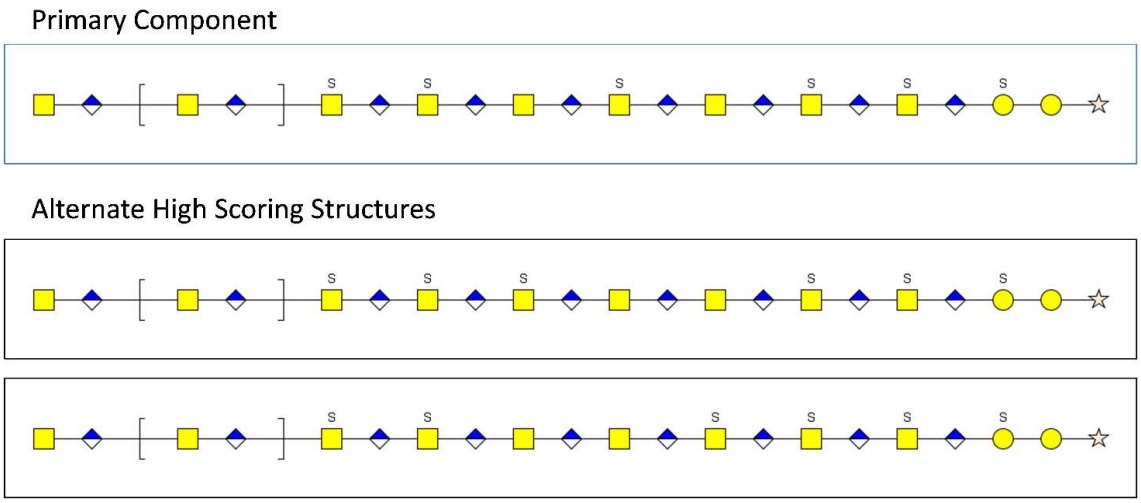


Figure 2.4. The highest scoring structure assigned to the all bikunin compositions (except d35-7S) provided, where the bracketed region is a variable stretch of unmodified disaccharides is outlined in blue. Two alternative structures are also frequently observed and outlined in black. The structures appear in the top 5 highest-scoring candidates for all compositions. For chain length dp43 (both $5SO_3$ and $6SO_3$), the highest score is tied amongst all three structures. Diagnostic fragments to confidently differentiate between these differences is absent.

In addition to matching previously reported structures, a closer examination of other high-scoring candidate structures among samples shows a consistent motif across compositions. Additional structural motifs shown in Figure 2.4 consistently score within the top 5 structures of the genetic algorithm. These alternate structures are ones consisting of similar f_1 and f_2 scores and but have low intensity values for some of their fragment matches (affecting the value of f_3). The high degree of similarity between the primary component identified in literature and the alternate structures may be a result of A) our scoring method being favored towards reducing end fragments, B) assigning low intensity

noise peaks as glycosidic fragments or C) the possibility of a mixture containing some minor components.

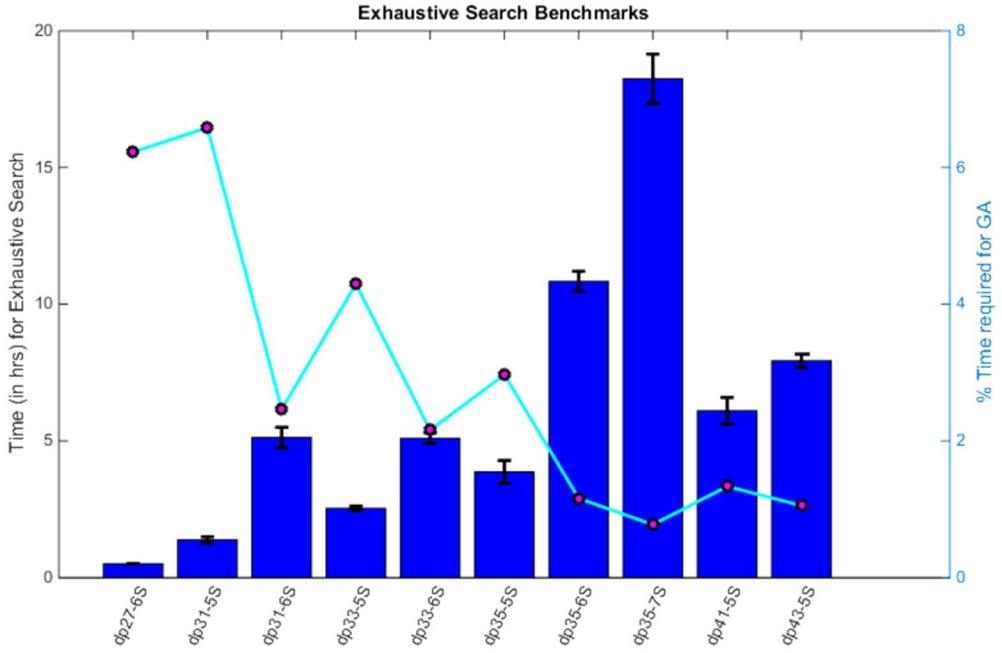


Figure 2.5. Speed comparison between the genetic algorithm and exhaustive search method. The bar graph shows the amount of time in hours (left y-axis) it requires for a standard desktop PC (2.4 GHz processor, 4GB ram) to exhaustively search through all possible combinations of a specific composition. The line plot shows the percentage of time (right y-axis) that is required for the genetic algorithm to arrive at the correct answer. Overall search space is reduced dramatically as the number of permutations per composition increases.

The speed of analysis between using the genetic algorithm versus the exhaustive search of every possible permutation of a composition is shown in Figure 2.5. Here we see that the genetic algorithm has found the correct answer within a small fraction of the time (0.9-2.5% on average) required to examine every possible structure with the assumption that sulfation only occurs on the 4-O position of the N-acetylgalactosamine. Decrease in

search time is primarily due to a reduction in the frequency in which unlikely features are eliminated from the genetic algorithm gene pool. As reported [41], bikunin's sulfation occurs near the reducing end. Isomeric structures that contain sulfate groups in the non-reducing end ranked lowest in the scoring process, resulting in rapid elimination of a test structure and all structures of similar sulfation patterns with one single iteration. A greater number of iterations were spent refining high-scoring structures once poorly scored structures have been eliminated from consideration. The algorithm is designed to rerun the entire genetic process from scratch multiple times in order to avoid plateauing at local maxima. Convergence upon the same highest scoring structure 5 times was the baseline criterion for an acceptable structural assignment. The repetition number is a user- adjustable parameter, as well.

Of particular significance, the efficiency of this approach is found to increase as the total number of permutations increases. For a pure sample, only a single structure can be assigned to the MS^2 spectrum, but the number of structures with drastically different modification patterns increases with respect to chain length. An increase in chain length also increases the number of GAG structures that could potentially share a feature not observed in the MS^2. Structures containing these features drop out of the algorithm as possible options once a single structure of that particular type is scored.

Calculations shown here are run on a 2.4 GHz dual-core processor with 4GB of RAM, a standard laptop or desktop computer. Speed of calculations can increase with more powerful processors such as a GPU workstation or computer cluster. It's important to note that the genetic algorithm in MATLAB is operated with separate function calls at each step of the algorithm's cycle. Parallelization of these function calls is particularly attractive for

samples of higher chain length and, in theory, could make spectra interpretation no longer the bottleneck for structural elucidation of GAGs. Additional GAG structures determined using this genetic-algorithm based GAG analysis software have been reported [53].

2.5 CONCLUSIONS

The software performance is limited by two factors: 1) the quality of the MS^2 data and 2) the specificity of the fitness function. The former limitation can be reduced by using a high-performance instrument such as FTICR or Orbitrap mass spectrometers. Some fragment mass values differ by less than 1 Da, increasing the possibility of ambiguity in low performance instruments. High resolution mass spectra with single digit or lower ppm mass error minimize margins for incorrect assignment. Acquisition condition must also be optimized for glycan fragmentation and ideally limits production of confounding fragments such as SO_3 loss or internal cleavages.

The latter factor, specificity of the fitness function in the genetic algorithm, is one that can be fine-tuned to GAG analysis by tandem mass spectrometry. The fitness function presented in this paper is simple, arbitrary and based on the basics of glycan analysis. This approach works for the examples selected here because only glycosidic bond cleavage was assigned. Higher level structure analysis based on cross-ring cleavages requires a more sophisticated fitness function. A more complete and non-arbitrary scoring algorithm is being developed that assigns statistical weights and importance factors to various fragment peaks. Additional, peak intensity, while not considered heavily in this iteration of the code, can also signify important characteristics in GAG structure. Details for creating an optimized scoring algorithm will be discussed in future work.

Peak picking for GAG fragmentation is not discussed in this paper but is an important consideration moving forward. Bikunin fragment peaks were selected by the SNAP algorithm using averagine and manually validated; this approach is practical for lowly sulfated samples but averagine is insufficiently for highly sulfated compounds due to contributions of sulfur to the A+2 isotope peak. A fully-automated and GAG-specific peak picking system is current in development.

The software is applicable for GAGs that are both lowly sulfated such as bikunin and moderate and highly sulfated samples for both CS/DS and HS/Hp samples. Short chain HS with more than one SO_3 modification per disaccharide and long chain chondroitin sulfate such as decorin with approximate 1 SO_3 per disaccharide have been determined using our software [52-53].

The uronic sugar stereochemistry is a variable modification in GAGs that is difficult to observe using just mass spectrometry. EDD data of heparin and heparan sulfate GAGs has produced a small subset of diagnostic fragments capable of distinguishing between glucuronic and iduronic acid epimers [22]. Chemometric applications has yielded a diagnostic fragment ratio that can definitively determine the C_5 stereochemistry [54]. Application of this ratio can be integrated into the software after basic structural features have been assigned using the approach presented here.

2.6 REFERENCES

1. Xie, B.; Costello, C. E., Carbohydrate Structure Determination by Mass Spectrometry. *Carbohydrate Chemistry, Biology and Medical Applications* **2008**, 29-57.
2. Gandhi, N. S.; Mancera, R. L., The Structure of Glycosaminoglycans and their Interactions with Proteins. *Chemical Biology & Drug Design* **2008,** *72* (6), 455-482.
3. Rabenstein, D. L., Heparin and heparan sulfate: structure and function. *Natural Product Reports* **2002,** *19* (3), 312-331.
4. Ohtsubo, K.; Marth, J. D., Glycosylation in cellular mechanisms of health and disease. *Cell* **2006,** *126* (5), 855-867.
5. Zhao, Y. J.; Singh, A.; Li, L. Y.; Linhardt, R. J.; Xu, Y. M.; Liu, J.; Woods, R. J.; Amster, I. J., Investigating changes in the gas-phase conformation of Antithrombin III upon binding of Arixtra using traveling wave ion mobility spectrometry (TWIMS). *Analyst* **2015,** *14* (20), 6980-6989.
6. Zhao, Y. J.; Singh, A.; Xu, Y. M.; Zong, C. L.; Zhang, F. M.; Boons, G. J.; Liu, J.; Linhardt, R. J.; Woods, R. J.; Amster, I. J., Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (1), 96-109.
7. Thanawiroon, C.; Rice, K. G.; Toida, T.; Linhardt, R. J., Liquid chromatography/mass spectrometry sequencing approach for highly sulfated heparin-derived oligosaccharides. *Journal of Biological Chemistry* **2004,** *279* (4), 2608-2615.
8. Jones, C. J.; Beni, S.; Limtiaco, J. F. K.; Langeslay, D. J.; Larive, C. K., Heparin Characterization: Challenges and Solutions. *Annual Review of Analytical Chemistry, Vol 4* **2011,** *4*, 439-465.
9. Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005,** *2* (9), 667-675.

10. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011,** *10* (4), 1794-1805.

11. Chi, L. L.; Amster, J.; Linhardt, R. J., Mass spectrometry for the analysis of highly charged sulfated carbohydrates. *Current Analytical Chemistry* **2005,** *1* (3), 223-240.

12. Cooper, C. A.; Gasteiger, E.; Packer, N. H., GlycoMod - A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **2001,** *1* (2), 340-349.

13. Kailemia, M. J.; Patel, A. B.; Johnson, D. T.; Li, L. Y.; Linhardt, R. J.; Amster, I. J., Differentiating chondroitin sulfate glycosaminoglycans using collision-induced dissociation; uronic acid cross-ring diagnostic fragments in a single stage of tandem mass spectrometry. *European Journal of Mass Spectrometry* **2015,** *21* (3), 275-285.

14. Flangea, C.; Serb, A. F.; Schiopu, C.; Tudor, S.; Sisu, E.; Seidler, D. G.; Zamfir, A. D., Discrimination of GalNAc (4S/6S) sulfation sites in chondroitin sulfate disaccharides by chip-based nanoelectrospray multistage mass spectrometry. *Central European Journal of Chemistry* **2009,** *7* (4), 752-759.

15. Huang, R. R.; Pomin, V. H.; Sharp, J. S., LC-MS (n) Analysis of Isomeric Chondroitin Sulfate Oligosaccharides Using a Chemical Derivatization Strategy. *Journal of the American Society for Mass Spectrometry* **2011,** *22* (9), 1577-1587.

16. Leach, F. E.; Xiao, Z. P.; Laremore, T. N.; Linhardt, R. J.; Amster, I. J., Electron detachment dissociation and infrared multiphoton dissociation of heparin tetrasaccharides. *Int. J. Mass Spectrom.* **2011,** *308* (2-3), 253-259.

17. Bin Oh, H.; Leach, F. E.; Arungundram, S.; Al-Mafraji, K.; Venot, A.; Boons, G. J.; Amster, I. J., Multivariate Analysis of Electron Detachment Dissociation and Infrared Multiphoton Dissociation Mass Spectra of Heparan Sulfate Tetrasaccharides Differing Only in Hexuronic acid Stereochemistry. *Journal of the American Society for Mass Spectrometry* **2011,** *22* (3), 582-590.

18. Wolff, J. J.; Laremore, T. N.; Leach, F. E.; Linhardt, R. J.; Amster, I. J., Electron capture dissociation, electron detachment dissociation and infrared multiphoton

dissociation of sucrose octasulfate. *European Journal of Mass Spectrometry* **2009,** *15* (2), 275-281.

19. Wolff, J. J.; Laremore, T. N.; Busch, A. M.; Linhardt, R. J.; Amster, I. J., Influence of charge state and sodium cationization on the electron detachment dissociation and infrared multiphoton dissociation of glycosaminoglycan oligosaccharides. *Journal of the American Society for Mass Spectrometry* **2008,** *19* (6), 790-798.

20. Leach, F. E.; Ly, M.; Laremore, T. N.; Wolff, J. J.; Perlow, J.; Linhardt, R. J.; Amster, I. J., Hexuronic Acid Stereochemistry Determination in Chondroitin Sulfate Glycosaminoglycan Oligosaccharides by Electron Detachment Dissociation. *Journal of the American Society for Mass Spectrometry* **2012,** *23* (9), 1488-1497.

21. Leach, F. E.; Wolff, J. J.; Laremore, T. N.; Linhardt, R. J.; Amster, I. J., Evaluation of the experimental parameters which control electron detachment dissociation, and their effect on the fragmentation efficiency of glycosaminoglycan carbohydrates. *International Journal of Mass Spectrometry* **2008,** *276* (2-3), 110-115.

22. Wolff, J. J.; Chi, L. L.; Linhardt, R. J.; Amster, I. J., Distinguishing glucuronic from iduronic acid in glycosaminoglycan tetrasaccharides by using electron detachment dissociation. *Analytical Chemistry* **2007,** *79* (5), 2015-2022.

23. Wolff, J. J.; Laremore, T. N.; Aslam, H.; Linhardt, R. J.; Amster, I. J., Electron-Induced Dissociation of Glycosaminoglycan Tetrasaccharides. *Journal of the American Society for Mass Spectrometry* **2008,** *19* (10), 1449-1458.

24. Wolff, J. J.; Laremore, T. N.; Busch, A. M.; Linhardt, R. J.; Amster, I. J., Electron detachment dissociation of dermatan sulfate oligosaccharides. *Journal of the American Society for Mass Spectrometry* **2008,** *19* (2), 294-304.

25. Huang, Y.; Yu, X.; Mao, Y.; Costello, C. E.; Zaia, J.; Lin, C., De Novo Sequencing of Heparan Sulfate Oligosaccharides by Electron-Activated Dissociation. *Analytical Chemistry* **2013,** *85* (24), 11979-11986.

26. Leach, F. E.; Riley, N. M.; Westphall, M. S.; Coon, J. J.; Amster, I. J., Negative Electron Transfer Dissociation Sequencing of Increasingly Sulfated Glycosaminoglycan Oligosaccharides on an Orbitrap Mass Spectrometer. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (9), 1844-1854.

27. Wolff, J. J.; Leach, F. E.; Laremore, T. N.; Kaplan, D. A.; Easterling, M. L.; Linhardt, R. J.; Amster, I. J., Negative Electron Transfer Dissociation of Glycosaminoglycans. *Analytical Chemistry* **2010,** *82* (9), 3460-3466.

28. Wolff, J. J.; Amster, I. J.; Chi, L.; Linhardt, R. J., Electron detachment dissociation of glycosaminoglycan tetrasaccharides. *Journal of the American Society for Mass Spectrometry* **2007,** *18* (2), 234-244.

29. Domon, B.; Costello, C. E., A SYSTEMATIC NOMENCLATURE FOR CARBOHYDRATE FRAGMENTATIONS IN FAB-MS MS SPECTRA OF GLYCOCONJUGATES. *Glycoconjugate J.* **1988,** *5* (4), 397-409.

30. Kailemia, M. J.; Li, L. Y.; Ly, M.; Linhardt, R. J.; Amster, I. J., Complete Mass Spectral Characterization of a Synthetic Ultralow-Molecular-Weight Heparin Using Collision-Induced Dissociation. *Analytical Chemistry* **2012,** *84* (13), 5475-5478.

31. Kailemia, M. J.; Ruhaak, L. R.; Lebrilla, C. B.; Amster, I. J., Oligosaccharide Analysis by Mass Spectrometry: A Review of Recent Developments. *Analytical Chemistry* **2014,** *86* (1), 196-212.

32. Zaia, J.; Costello, C. E., Tandem mass Spectrometry of sulfated heparin-like glycosaminoglycan oligosaccharides. *Analytical Chemistry* **2003,** *75* (10), 2445-2455.

33. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A., De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* **1999,** *6* (3-4), 327-342.

34. Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **2003,** *17* (20), 2337-2342.

35. Taylor, J. A.; Johnson, R. S., Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry* **2001,** *73* (11), 2594-2604.

36. Campbell, M. P.; Hayes, C. A.; Struwe, W. B.; Wilkins, M. R.; Aoki-Kinoshita, K. F.; Harvey, D. J.; Rudd, P. M.; Kolarich, D.; Lisacek, F.; Karlsson, N. G.; Packer, N. H., UniCarbKB: Putting the pieces together for glycomics research. *Proteomics* **2011,** *11* (21), 4117-4121.

37. Maxwell, E.; Tan, Y.; Tan, Y.; Hu, H.; Benson, G.; Aizikov, K.; Conley, S.; Staples, G. O.; Slysz, G. W.; Smith, R. D.; Zaia, J., GlycReSoft: A Software Package for Automated Recognition of Glycans from LC/MS Data. *Plos One* **2012,** *7* (9).

38. Saad, O. M.; Leary, J. A., Heparin sequencing using enzymatic digestion and ESI-MSn with HOST: A heparin/HS oligosaccharide sequencing tool. *Analytical Chemistry* **2005,** *77* (18), 5902-5911.

39. Chiu, Y. L.; Huang, R. R.; Orlando, R.; Sharp, J. S., GAG-ID: Heparan Sulfate (HS) and Heparin Glycosaminoglycan High-Throughput Identification Software. *Mol. Cell. Proteomics* **2015,** *14* (6), 1720-1730.

40. Hu, H.; Huang, Y.; Mao, Y.; Yu, X.; Xu, Y. M.; Liu, J.; Zong, C. L.; Boons, G. J.; Lin, C.; Xia, Y.; Zaia, J., A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra. *Mol. Cell. Proteomics* **2014,** *13* (9), 2490-2502.

41. Ly, M.; Leach, F. E., III; Laremore, T. N.; Toida, T.; Amster, I. J.; Linhardt, R. J., The proteoglycan bikunin has a defined sequence. *Nature Chemical Biology* **2011,** *7* (11), 827-833.

42. Baeck, T.; Schwefel, H.-P., An Overview of Evolutionary Algorithms for Parameter Optimization. *Evolutionary Computation* **1993,** *1* (1), 1-23.

43. Fogel, L. J.; Owens, A. J.; Walsh, M. J., Artificial intelligence through a simulation of evolution. *Proceedings of the Second Cybernetic Sciences Symposium: Biophysics and cybernetic systems* **1965**, 131-155.

44. Forrest, S., GENETIC ALGORITHMS - PRINCIPLES OF NATURAL-SELECTION APPLIED TO COMPUTATION. *Science* **1993,** *261* (5123), 872-878.

45. Han, L.; Costello, C. E., Mass spectrometry of glycans. *Biochemistry-Moscow* **2013,** *78* (7), 710-720.

46. Kilgour, D. P. A.; Neal, M. J.; Soulby, A. J.; O'Connor, P. B., Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm. *Rapid Communications in Mass Spectrometry* **2013,** *27* (17), 1977-1982.

47. Das, S.; Suganthan, P. N., Differential Evolution: A Survey of the State-of-the-Art. *IEEE Trans. Evol. Comput.* **2011,** *15* (1), 4-31.

48. Knowles, J. D.; Corne, D. W., Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy. *Evolutionary Computation* **2000,** *8* (2), 149-172.

49. Phillips, S. J.; Anderson, R. P.; Schapire, R. E., Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **2006,** *190* (3-4), 231-259.

50. Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; Church, G. M., Systematic determination of genetic network architecture. *Nature Genetics* **1999,** *22* (3), 281-285.

51. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins-Structure Function and Genetics* **2003,** *52* (4), 609-623.

52. Yu, Y. L.; Duan, J. N.; Leach, F. E.; Toida, T.; Higashi, K.; Zhang, H.; Zhang, F. M.; Amster, I. J.; Linhardt, R. J., Sequencing the Dermatan Sulfate Chain of Decorin. *J. Am. Chem. Soc.* **2017,** *139* (46), 16986-16995.

53. Singh, A.; Kett, W. C.; Severin, I. C.; Agyekum, I.; Duan, J. N.; Amster, I. J.; Proudfoot, A. E. I.; Coombe, D. R.; Woods, R. J., The Interaction of Heparin Tetrasaccharides with Chemokine CCL5 Is Modulated by Sulfation Pattern and pH. *Journal of Biological Chemistry* **2015,** *290* (25), 15421-15436.

54. Agyekum, I.; Patel, A. B.; Zong, C. L.; Boons, G. J.; Amster, I. J., Assignment of hexuronic acid stereochemistry in synthetic heparan sulfate tetrasaccharides with 2-O-sulfo uronic acids using electron detachment dissociation. *Int. J. Mass Spectrom.* **2015,** *390*, 163-169.

CHAPTER 3

A STRUCTURAL IDENTIFICATION PARADIGM FOR CHARACTERIZING GLYCOSAMINOGLYCANS FROM TANDEM MASS SPECTROMETRY*

Duan, J.N.; Pepi, L.E.; Amster, I. J. **2018.** *Preparing for submission to Journal of the American Society of Mass Spectrometry.

3.1 ABSTRACT

The role of glycosaminoglycans (GAGs) in major biological functions are numerous and diverse, yet structural characterization of them by mass spectrometric techniques proves to be challenging. Characterization of GAG structure from tandem mass spectrometry is a tedious and time-consuming process but one that can be automated in a database-independent, high-throughput fashion through the assistance of software implementing a genetic algorithm. The present work presents how this data is interpreted by the software, specifically addressing the form of a scoring algorithm. The significance of glycosidic and cross-ring fragment ions and the implications that specific fragments provide for assigning the positions of modifications are discussed. The scoring algorithm is tested for statistical merit using the widely accepted expectation value as the criterion for quality. Using MS/MS data for well-characterized standards, this scoring approach is shown to assign the correct structure, with a low likelihood (1 in 10^{12} chance) that the assigned structure matches the data due to random chance.

3.2 INTRODUCTION

Glycosaminoglycans (GAGs) are linear, polydisperse carbohydrates that are ubiquitous among living cells, and are responsible for a multitude of biological interactions including cell signaling, energy generation, protein binding conformation changes and molecular recognition [1-4]. Structurally, GAGs are composed of a repeating linear disaccharide backbone of a uronic sugar and amino sugar residue. Structural differentiation occurs based on three primary forms of modifications: *O*-sulfation, *N*-deacetylation/sulfation and uronic sugar epimerization. Recent studies suggest that

patterns of sulfation has profound effect on protein binding[5-6]. Moreover, analysis of GAGs released from proteoglycans bikunin and decorin suggest that naturally occurring glycans may contain a conserved sulfation motif that is independent of chain length [7-8]. Modifications can be localized to specific glycan residues and further designated to specific positions with the respective use of glycosidic and cross-ring fragmentation that arise from ion activation using collisions, photodissociation, or electron activation [9-19].

The importance of structural modifications to the biological activity of GAGs cannot be overstated, and considerable effort has been made to develop new approaches for sequencing this class of molecules. The assignment of structures for GAGs require a *de novo* approach in which the mass spectra must contain sufficient information to localize all the relevant features. There are several methods that can be used to produce information-rich tandem mass spectra, using collisions, photodissociation, or electron-based activation methods. The density of fragment ions in these tandem mass spectra makes it difficult to interpret and assign structures in a high-throughput manner without assistance from specialized software. Manual interpretation of tandem MS of GAGs is possible but requires both a great deal of time and expertise. While the interpretation of known structures can be significantly improved with mass calculation tools such as GlycoWorkBench [20], the feasibility of interpreting unknown GAG structures diminishes greatly with increasing degree-of-polymerization (dp), as the number of permutations scales as *n*-choose-*k*, where *n* is number of possible modification sites (up to 4 per disaccharide) and *k* is number of modifications. Previous work from our laboratory proposes software for a database-independent method of automated GAG tandem MS interpretation using a genetic algorithm. It should be noted that our software is noticeably

different compared to databased linked methods that have been used by Zaia et al. which connects to a database first and uses expected fragments from said database to determine structure. Our method hinges on being able to assign composition first and foremost from accurate mass measurement of the MS^1, and from this step generate fragments *in silico* before optimizing structural possibilities with a genetic algorithm. Optimization is reliant on a survival of the fittest mechanism (and therefore a scoring criterion) for what theoretical structures match the "fitness" of the experimental mass spectrum; this manuscript explains in detail the software features and scoring methods that enable this approach.

3.3 EXPERIMENTAL METHODS

Mass Spectrometry. Electron detachment dissociation (EDD) experiments were performed using a 9.4T Bruker Apex Ultra QeFTMS (Billerica, MA), with an indirectly heated hollow cathode (HeatWave, Watsonville, CA) for generating EDD electrons. The solutions were ionized using flow nanoelectrospray (pulled fused silica tip FS360-75-15-N-20) Solutions were made at a concentration of 0.2 mg/mL in 50:50 methanol:H_2O. Hexasaccharide solutions were injected at a rate of 25uL/h. The solutions were run in negative ion mode. Precursor ions were selected in the external quadrupole accumulated for 4 seconds in an RF only hexapole before injection into the Fourier transform-ion cyclotron resonance mass spectrometer (FT-ICR MS). Precursor isolation was refined by using in-cell isolation with coherent harmonic excitation frequency (CHEF). The isolation power of the CHEF event was 20%. For irradiation of electrons the cathode heater was set to 1.6 A and the bias was set to -19 V. The extraction lens was set to -19.2±0.2 V. Ions

were irradiated for 1 s. 64 acquisitions were signal averaged per mass spectrum. Internal calibration produced a mass accuracy of 5 ppm.

Software. MS^1 analysis of parent ion mass is performed using a composition assignment software module written in the MATLAB coding environment. Monoisotopic peaks and charge states are acquired from Bruker DataAnalysis (using the "FTMS" peak picking method) and deconvoluted to a neutral mass. A composition is derived from one or more neutral mass(es) by searching a data matrix of possible chain lengths, degrees of sulfation, deacetylation, and sodium/hydrogen exchange. The user input also includes the possibility of reducing end modifications, and nonreducing ends that can terminate in unsaturated uronic acids, as is common in enzymatically produced GAG oligomers. Theoretical neutral masses in the spreadsheet are compared against user specified masses with a user-defined mass tolerance. The sequences that match are then used for performing the MS^2 analysis.

For MS^2 analysis, the software uses a genetic algorithm search heuristic alongside binary vector representation of glycan structures where on-bits denote an occupied site of SO_3 and N-acetylation modifications. The first step generates two glycan structures at random that fit the expected composition (*initialization step*) and then proceeds to "breed" these structures into a new generation of candidates (*crossover step*). The primary three steps (crossover, mutation and fitness) are iterated until the maximum fitness value does not change after numerous cycles. The number of iterations required before termination of the algorithm can be defined by the user but is defaulted at a value of 3. The structure(s) containing the highest scores are then examined using additional data interpretation tools that assign fragment peak masses alongside their charge, intensity and mass error (in ppm).

Software is compatible for standard desktop computers, with a minimum requirement of MATLAB R2014 coding environment or newer, 2.4 GHz processor and 4GB RAM. Customization specific to GAG family (CS/DS, Hp/HS and Arixtra) is available using specific functions for determination of fragment masses. Additional parameters are adjustable from the command line if desired (maximum charge state of fragments, ppm mass error tolerance, maximum possible Na-H exchange, neutral loss considerations for H_2O and CO_2). MATLAB source code is available upon request.

3.4 RESULTS AND DISCUSSION

Software Architecture

Previous work from our laboratory discussed a high-throughput method for characterizing GAG structure of unknown samples from tandem MS using a database independent method that generates theoretical fragments *in-silico* and then proceeds to optimize potential structures using a genetic algorithm. Figure 3.1 shows the workflow that is used: a user provides tandem MS data in the form of a two-column comma-separate value (.csv) file with masses (monoisotopic, or all isotopes? Deconvoluted or not?) sorted ascending in column 1 and intensities (relative or arbitrary units) in column 2 as well as additional inputs of the parent ion mass and charge of the precursor ion. An independent module calculates the composition -degree of polymerization, number of SO_3 groups, number of acetyl groups - of the sample within a user defined ppm tolerance window. Once completed, the software automatically employs a genetic algorithm optimization model to determine the most likely structure(s) based on the data provided.

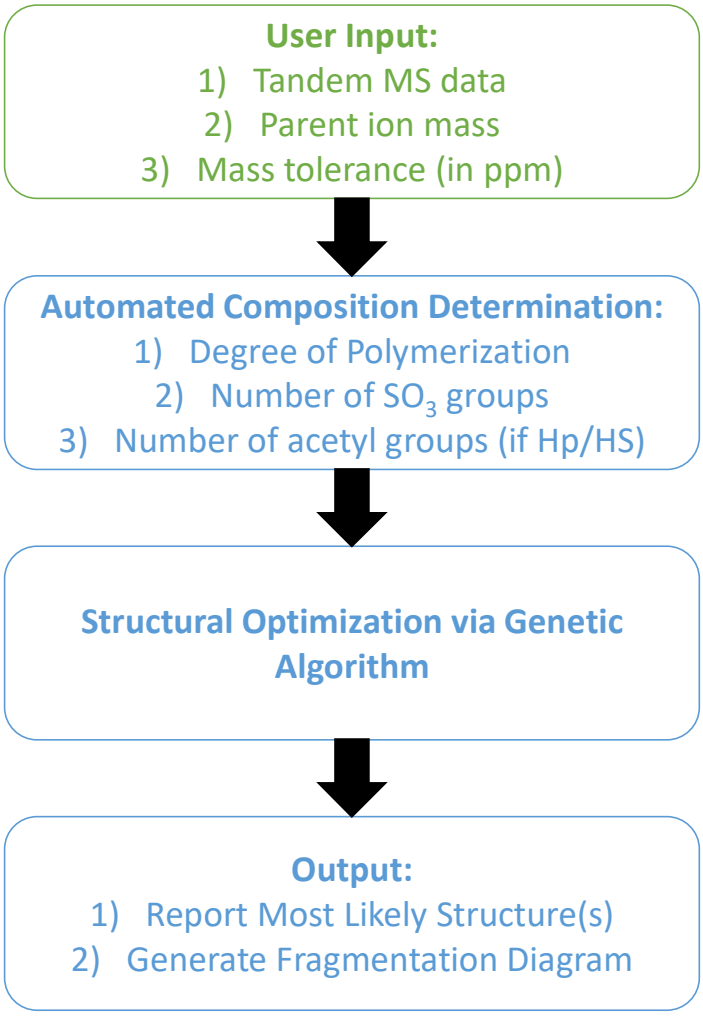


Figure 3.1. The standard workflow for our current GAG identification software. The user provides the information in the green box. Blue boxes are connected module, fully automated and require no user supervision.

The software generates a list of theoretical structures of appropriate composition to test against experimental data, and scores the structures based on their probability for producing the observed fragment ions. A prior publication has shown how we can represent GAGs as computational vectors that are practical for genetic algorithms, Figure 3.2. Each iteration of the genetic algorithm attempts to improve upon the list of theoretical structures while eliminating low scoring structures from consideration. This paper focuses

on the comprehensive scoring model that can be applied to any GAG family (heparin, heparan sulfate, chondroitin sulfate, dermatan sulfate) and any chain length. The focus is on the methodology used to compare computer generated theoretical structures against experimental data and to show the statistical merit of the scoring paradigm.

While the earlier proof-of-principle example using bikunin GAGs only examined glycosidic fragments [7], most GAG samples are far more structurally diverse. Modification heterogeneity and the potential for multiple sulfo-modifications to occur on the same sugar residue dictate that we not only incorporate cross-ring fragmentation into our scoring model but also focus in on structurally diagnostic fragments or fragment sets that can yield unambiguous assignments of modification position.

Scoring Algorithm

The fundamental step that allows this software to be independent of databases is that it considers all possible isomers, constrained by GAG family and by composition (which constrains dp, degree of sulfation, and the number of acetylated amine groups). In order to allow all isomers to be considered, without having to score every possibility, the genetic algorithm optimizes the match of structures to the experimental data. Using this optimization tool allows scoring on a small subset (1% or less) of all possible permutations of a given composition while finding the correct structure. Within this optimization step is a series of comparisons between theoretical structures and the experimental data – each iteration of the genetic algorithm attempts to find a more closely matched theoretical structure based on its interpretation of experimental data; the *fitness score* of a theoretical

structure is the value used as a measure of closeness. Hence, the paradigm for how the software scores theoretical structures must be discussed.

A 2-step system is utilized in which an initial score is given to structures based on glycosidic fragments alone and then a refined score is assigned based on the number of cross-ring fragments that can be used to unambiguously define a specific modification location. The first step determines the number and type of modifications that are located on a particular residue. For chondroitin sulfate and dermatan sulfate (CS/DS) GAG families, we consider the possibility of *2-O* sulfation on the hexuronic residue and *4-O* and *6-O* sulfation on the N-acetyl galactosamine. Optionally, more uncommon CS modifications such as *N*-sulfation or a free amine group can be considered, in which case a new bit is introduced (Figure 3.2). For heparin and heparan sulfate (Hp/HS) GAG families, we consider the possibility of 2-O sulfation on hexuronic residues and *3-O*, *6-O* and *N*- sulfation on glucosamine residues. Additionally, de-acetylated glucosamines are also an optional feature.

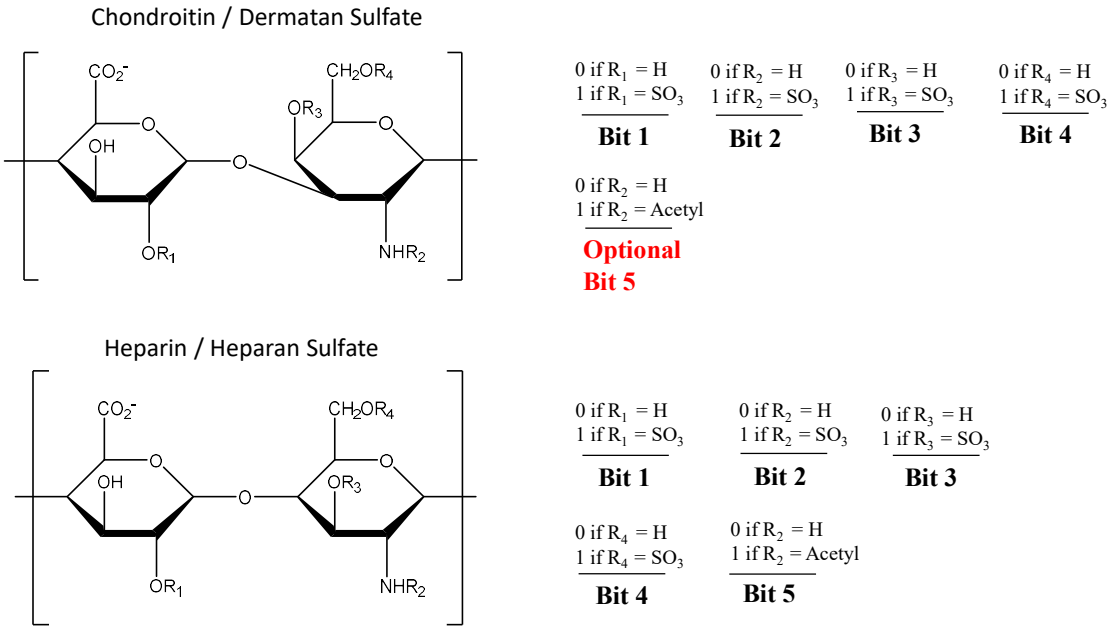


Figure 3.2. Bit-wise matrix representation of disaccharide CS/DS and Hp/HS GAG families used in software for rapid analysis via a genetic algorithm. Multiple disaccharide units can be combined to fit the appropriate chain length and composition.

Glycosidic fragments are used as a confining mechanism: for each sugar residue, that sugar residue is assigned a score of 0 to 4 (called GlycScore or *SG*), ranked based on the number of glycosidic fragments that surround the sugar residue as well as what information these fragments provide. Figure 3.3 provides the criterion for assigning a score to a residue: a GlycScore of 4 indicates that glycosidic cleavages are observed pointing directly to the specific residue and come from both the reducing and non-reducing ends. On the other hand, a low-value GlycScore indicates only a few or no fragments that are supportive of a particular residue. Assuming a high quality of tandem MS data and information rich spectra, we should expect that the majority of glycosidic fragments are present. When using a genetic algorithm optimization method, theoretical structures that

lack a large fraction (>50%) of glycosidic fragment matches are evaluated as unlikely matches and other structures with similar characteristics are unlikely to be examined.

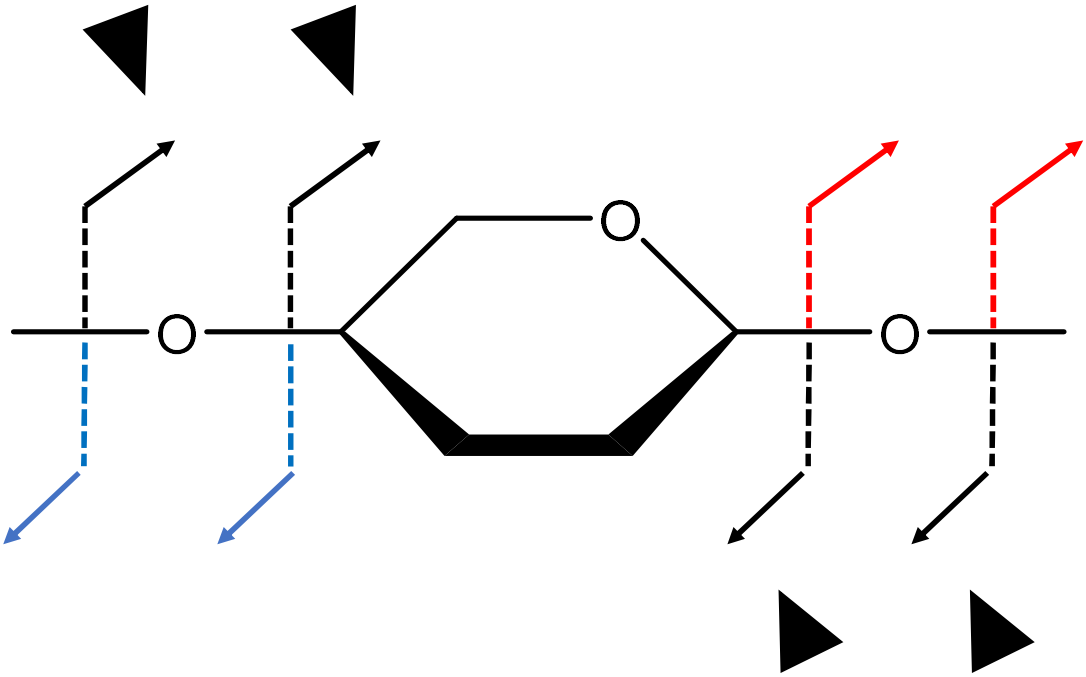
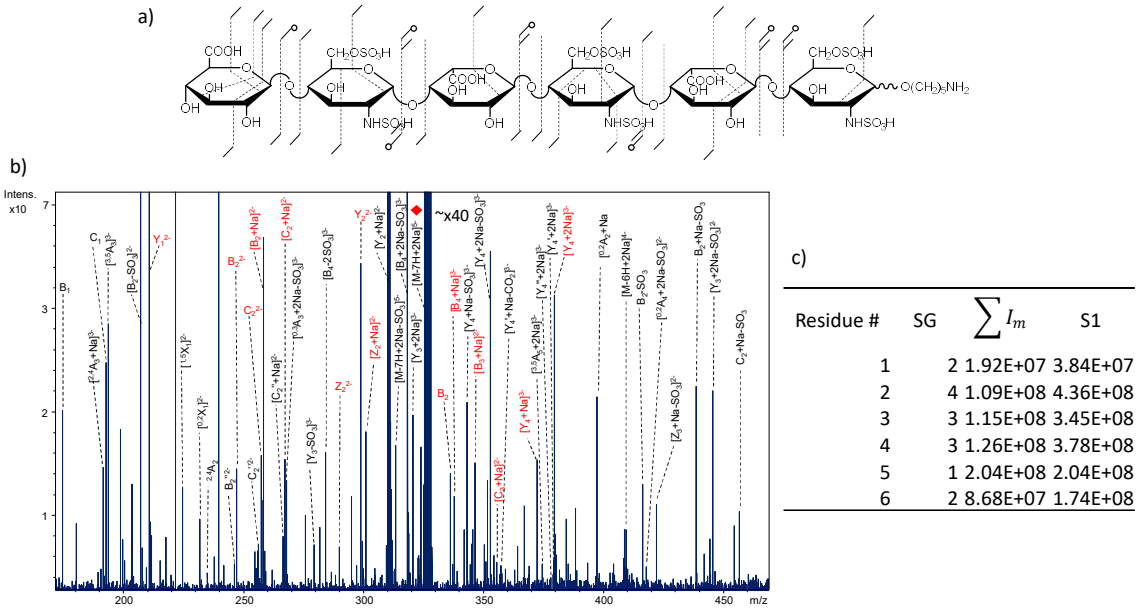


Figure 3.3. A simple paradigm for assigning GlycScore *(SG)* to a residue. Each residue's *SG* is concerned with the set of non-reducing and reducing end fragments that point to it. For residue *n*, B_n, C_n, $Y_{max-n+1}$ and $Z_{max-n+1}$ (shown with black arrows) are the fragments considered to be part of SG_n . Fragments in blue contribute to the *SG* of the residue adjacent from the non-reducing end. Likewise, red fragments contribute to the *SG* of the residue adjacent from the reducing end.

The ranking of structures and their respective step 1 score (*S1*) is not solely reliant on the GlycScore (*SG*) but also the intensities of fragment matches as shown in eq 1, where I_n is intensity of the 4 fragments (B_n, C_n, $Y_{max-n+1}$,$Z_{max-n+1}$) that point directly to the residue, *n*, in question:

$$(eq.1) \qquad S1_n = SG_n * \sum_{m=1}^{4} I_m$$

The GlycScore by itself will always yield an integer value and is prone to problematic artifacts if used as the sole measurement for ranking structures as multiple structures can have the exact same GlycScore if the number of glycosidic fragments matches are equal. The presence of SO_3 losses in the tandem MS exacerbates the problem as loss fragments can be matched as a loss-free glycosidic fragment when the genetic algorithm presents a theoretical structure that is under sulfated in a specific region compared to the true structure. The solution is to consider the intensity of matched glycosidic fragments (I_m) as a scaling factor to *SG*. Previous work on GAG tandem MS for multiple ion activation techniques include CID, EDD, and NETD show that the intensity of SO_3 loss peaks are typically lower in abundance than the non-loss peak [11-12, 16-18]. Figure 3.4 details and explains assigning scores *SG* and *S1* for synthetic hexasaccharide GAG with 6 sulfo-modifications.



Residue #	SG	$\sum I_m$	S1
1	2	1.92E+07	3.84E+07
2	4	1.09E+08	4.36E+08
3	3	1.15E+08	3.45E+08
4	3	1.26E+08	3.78E+08
5	1	2.04E+08	2.04E+08
6	2	8.68E+07	1.74E+08

Figure 3.4. **(a)** Annotated structure of synthetic hexasaccharide sample with 6 SO_3 groups; fragments were produced by electron detachment dissociation (EDD). **(b)** Annotated EDD spectrum of structure from 4a. **(c)** Calculated *SG, I_m* and *S1* values for each residue *n*. Note that in the table, the sum of all intensities for relevant fragment peaks is shown.

One challenge in assigning glycosidic fragments occurs when there are isobaric possibilities for product ions from symmetrically modified oligomers. This happens for c and z ions from a precursor with a delta-uronic acid at the non-reducing end and no derivatization of the reducing end. Isobaric fragment ions are difficult to assign either by manual interpretation or by automated glycan analysis – our software currently acknowledges isobaric ions as both forms of fragment ions as a clear, unambiguous method of distinguishing the two is not available. This problem is mitigated for GAG samples where the reducing end has been modified in some capacity to create an unambiguous mass difference. In the case of intact GAGs from proteoglycans, the linker region breaks the symmetry of the structure and eliminates the possibility of isobaric c/z ions [7-8].

The second step of our algorithm uses a logic-based decision tree (Figure 3.5) to determine the position of SO_3 modifications on sugar residues for the top scoring structures from step 1 (by default the top 3 structures). At this point in the program, the optimal structure is already exposed to a great deal of modification constraints based on glycosidic fragment matches. Therefore, it is possible to examine every glycan residue individually as there is a narrow selection of permutations based on the number of biologically-possible modifications available. For hexuronic acids, the presence of an SO_3 defaults to being at the 2-O position. For amino sugars, the code looks for diagnostic cross-ring fragments or fragment combinations that would be able to unambiguously assign SO_3 positions. If no cross-ring fragments are available on a residue, the code will leave the positions on that residue as ambiguous. On the other hand, in situations where multiple structures arise from conflicting cross-ring assignments, we rank structures based on the intensities of the diagnostic cross ring fragments.

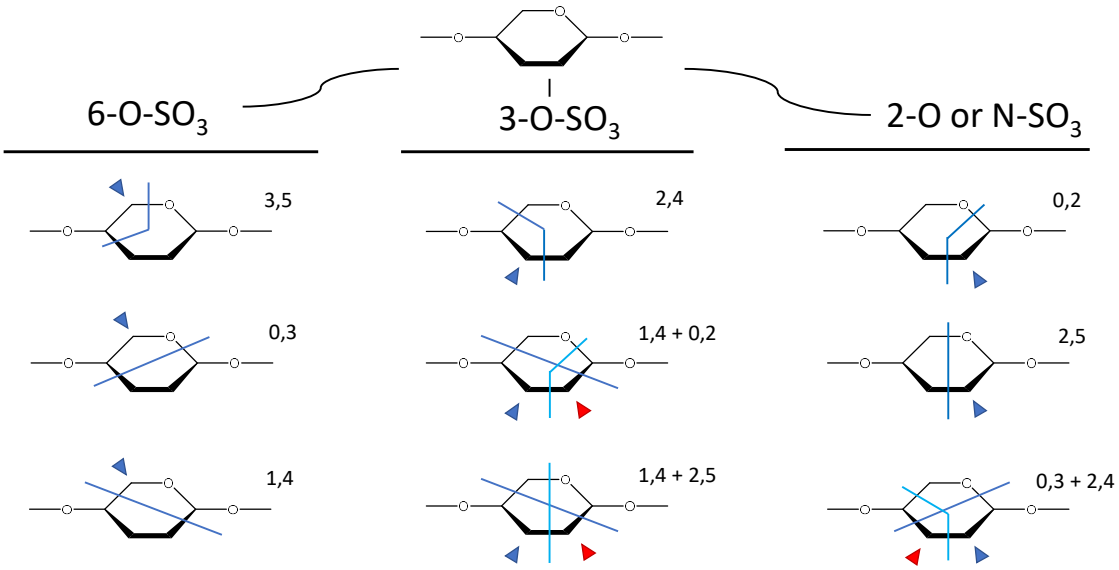


Figure 3.5. Once the number of SO_3 on a residue are confined to a specific residue, software searches for cross-ring fragments to determine position. In situations where a mixture might be present, structures are ranked based on the cumulative intensity of diagnostic fragments. Blue arrows indicate positions that can be assigned with a specific cross ring fragment or a combination of 2 cross-ring fragments with the assumption that they come from the same end (reducing or non-reducing). A red arrow indicates assignments that can be made if the two indicated cross-ring fragments are from different ends (i.e. an A fragment and an X fragment).

Intensity ratios between sets of fragments have been used to differentiate the presence of glucuronic versus iduronic acids in synthetic HS tetrasaccharides [21]. Furthermore, multivariate statistics have been used to identify diagnostic fragments for certain ion activation methods [11, 14, 22-24]. At the present level of knowledge, there is not sufficient knowledge to use these ratios as a definitive guide to determining all C-5 stereochemistry. The software is fully capable of applying an additional layer of analysis to determine the epimeric center once there is a fuller understanding of product ion intensities as a function of uronic acid stereochemistry. Much like the method proposed

in step 2, a residue-by-residue analysis of the uronic sugars could be applied as a 3rd step without a significant increase in analysis time, and this capability is envisioned in a future release.

Validation of Algorithm

In order to create a fully automated GAG interpretation software, the scoring algorithm must be 1) unsupervised and user-independent and 2) assign the correct glycan structure. Much of the glycan analysis currently published in literature uses software packages such as GlycoWorkBench [20] or similar GAG fragment calculation tools in combination with user intuition or experience to interpret the mass spectrum. Careful examination of peaks in the mass spectrum with manual supervision allows an expert to determine the likelihood of false positive while reaffirming structural features. This in-depth yet subjective form of glycan interpretation that relies on user expertise is difficult to automate and impractical for high-throughput analysis.

Assigning structure from tandem MS fragment peaks in an automated fashion requires some degree of assumption - for our software, we assume glycosidic and cross-ring fragments increase the validity of certain structural features and that enough of them will give the highest score to the most valid structure. However, the quality of the scoring system is judged not on its theoretical foundation but purely by it's ability to assign the correct sequence and differentiate it from incorrect ones. An objective approach to determining the quality of our scoring system involves taking a statistical approach: the likelihood that a score is given due to random chance can be determined as well by calculating its expectation value, a statistic that has been widely applied and accepted in

bioinformatics [25-27]. If x is the score of a particular spectrum $\boldsymbol{S}$, a survival function, *s(x),* for a discrete score probability distribution, *p(x)*, can be defined:

$$(eq.2) \quad s_{j(x)} = Pr(X > x) = \sum_{i=j(x)}^{\infty} p_i$$

Where Pr($X > x$) is defined as probability that the spectrum's score will be higher than score x due to random matching within a defined database, $\boldsymbol{D}$. For GAGs, the defined database is all possible permutations of the composition of the sample of spectrum $\boldsymbol{S}$. The expectation value *e(x)* can be interpreted as the number of GAG structures that would be expected to have scores of at least x.

$$(eq.3) \quad e_{j(x)} = n * s_{j(x)}$$

Where n is the number of sequences scored. The expectation value can be interpreted as follows: if a score x of expectation value $e(x) = y$, then one would have a score of at least that value for y number of times for every replicate experiment. A lower expectation value is therefore more ideal. For example, an *e(x)* of 0.001 suggests that an experiment must be replicated 1 thousand times before a score of x could be obtained by random chance.

This technique has been used previously for analysis of scoring system of peptide MS [28]while using a database search engine [29]; we apply the same fundamental principles and calculations for our scoring system. To examine our GAG scoring system with this method, *p(x)* is determined by constructing a frequency histogram of all GAG structure scores. We take the tandem MS of a pure, single component GAG for which we know the structure (Figure 3.6a) and score structures of appropriate composition against the experimental data. The structures being scored against are stochastically selected and not optimized with the genetic algorithm heuristic to prevent introduction of selection bias. Among these structures, we know that only one is considered "valid" while all others are

termed “stochastic”. The probability *p(x)* can be determined by normalizing the discrete frequency *f(x)* of a structure by the number of sequences scored *N*:

$$(eq.4) \qquad p_i = \frac{f_i}{N}$$

Figure 3.6b is the frequency histogram of 20,000 structures scored using a Monte Carlo sampling method for a synthesized HS hexasaccharide containing 8 SO_3 groups. The green asterisk in the histogram represents the score associated with the valid GAG structure. Visual inspection of the histogram shows that the right-most point of the histogram contains the structure of highest score, x^*, and is the valid structure for this data set. The confidence of this score increases with respect to the distance between x^* and all other scores (the scores of stochastic structures). Hence, a gap between x^* and the bulk majority of other scores is highly desired and, moreover, the difference between x^* and the next highest scoring structure is important for evaluating the algorithms ability to discriminate similar structures. It should also be noted that the frequency of the highest score x^* is the lowest in the set, implying that x^* is likely a unique value observed only when matched with the valid structure.

Given how our scoring system is partially dependent on intensity while also working under some assumptions regarding the expected fragmentation results from ion-activation of GAGs, the numerical difference between two scores is not an easily interpretable measurement of the degree of difference between two structures of those respective scores. Moreover, the individual score of any structure (valid or stochastic) has

little interpretability and does not serve as a good measure of fitness. A much more sophisticated estimate for confidence can be determined from the survival function (Figure 3.6c). Score x^* for our hexasaccharide has a value of $s(x^*) = 5.59E\text{-}16$; application of *eq.3* yields an $e(x^*) = 1.96E\text{-}12$. This expectation value indicates that an experiment would have to be repeated approximately 10^{12} times before a score of x^* would be matched to a structure due to random chance – a figure of merit that reflects positively on our GAG scoring algorithm.

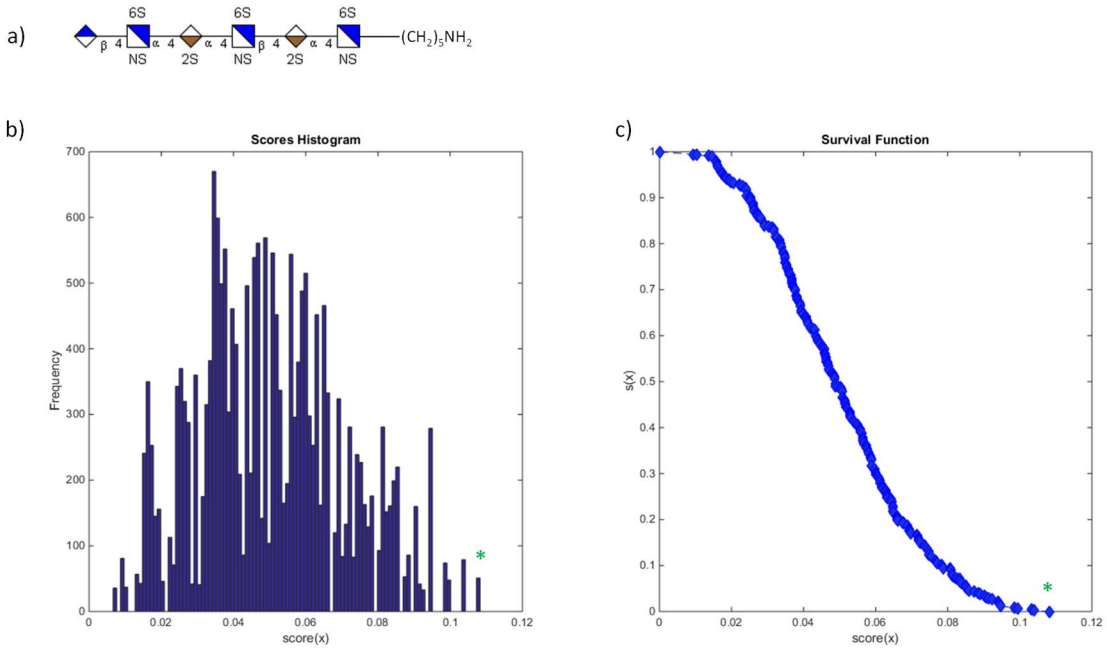


Figure 3.6. (a) The hexasaccharide structure used for calculating *e(x)* and *s(x)*. Tandem MS was performed using EDD. **(b)** A histogram of 20,000 structures scored using the Monte Carlo method. The * represents the score that is associated with the structure shown in figure 3.6a. Note that the scoring algorithm assigns the valid structure the highest score. **(c)** The survival function of the scoring algorithm plotted versus the score. An *e(x)* value of 1.96E-12 is calculated for the valid structure.

Additionally, histograms and survival function diagrams for other GAG compositions of both experimental and synthetically generated datasets for CS/DS and Hp/HS GAG families were also calculated and available in supplemental data. The expectation values for the scores of valid structures (x^*) using our scoring algorithm across various chain lengths, GAG families and different degrees of modification suggests that our method can be applied to a wide variety of GAG tandem MS.

Isotope Deconvolution

The averagine model ($C_{4.9384}$ $H_{7.7583}$ $N_{1.357701}$ $O_{1.4773}$ $S_{0.0417}$) has been a standard for protein isotope deconvolution in mass spectrometry but is severely limited in its applicability for GAG analysis. Highly sulfated GAGs and GAG tandem MS fragments exhibit an A+2 isotope peak with significantly greater intensity than that of averagine. Either the means for assessing the validity of isotope clusters in GAGs must change or a new model for predicting glycan isotope distributions should be developed. A dot product comparison of closeness, where X is the intensity of an isotope distribution, is fundamentally simple and accurate; modifying a straightforward comparison method could lead to unintentional artifacts within software.

$$(eq.\ 5) \qquad \theta = acos \frac{X_{theoretica} \cdot X_{experimental}}{\|X_{theoretical}\| * \|X_{experimental}\|}$$

Therefore, we propose an alternate means for performing deconvolution and determining monoisotopic peaks by starting with a variation in the average amounts of C, H, N, O and S to $C_{2.7710}$ $H_{4.3875}$ $N_{0.2309}$ $O_{3.2329}$ $S_{0.2309}$. This distribution is the molecular composition per 100 Da for a GAG containing exactly 1 SO_3 per disaccharide. Peaks

within a mass spectrum are examined against this distribution using the comparison method as described, with a default tolerance value ranging from 0-20 degrees being acceptable for θ. The range limit is adjustable by the user as the quality of the experimental isotopes can heavily impact its comparison with theoretical values.

Highly sulfated fragment ions can exhibit isotope distributions with abundances that deviate from our default condition by a θ value >20 degrees. In these situations, only a fraction of all monoisotopic peaks will be identified with a single iteration; multiple iterations where the number of SO_3 per disaccharide is incrementally expands the range of isotope distributions that are possible. The preprocessing step is thus done in layers, starting with 1 SO_3 group per disaccharide, extracting all matching isotope clusters, and then incrementing the number of SO_3 groups until a practical limit is reached (typically a maximum of 4 SO_3 groups per disaccharide).

3.5 CONCLUSIONS

While calculating expectation values, we were fortunate enough to have datasets with the necessary amount and type of fragments needed for structural characterization. However, the acquisition of data can be a limiting factor in determining structure. Much like any form of automated spectra interpretation, the success of the automated approach is highly reliant on the quality of the data: glycosidic fragments and structurally meaningful cross-ring fragments are necessary for complete structural characterization. A lack of any necessary pieces of information both increase the possibility of structural ambiguity and simultaneously increases processing times. With this software, glycosidic fragments are particularly important as they confine certain structural features and make cross-ring

fragments more readily interpretable. Although the likelihood of only observing cross-ring fragments without supporting glycosidic fragments is unlikely in tandem MS, this hierarchical system of structural interpretation would have difficulty interpreting a structure from cross-ring fragments alone.

Data preprocessing methods have not been discussed in this manuscript as there are various specialized considerations that must be made for GAGs. Deconvolution of GAGs using commercialized options such as Bruker DataAnalysis or Thermo XCalibur fail to capture all relevant monoisotopic peaks. This is because isotope distributions of GAG fragments deviate heavily from the typical proteomic averagine model, and more importantly, do not necessarily fall under a consistent chemical formula but instead change with respect to the number of SO_3 modifications. Isotope deconvolution thus needs its own set of specialized rules, a point we will discuss with more detail in a separate manuscript.

3.6 REFERENCES

1. Gandhi, N. S.; Mancera, R. L., The Structure of Glycosaminoglycans and their Interactions with Proteins. *Chemical Biology & Drug Design* **2008,** *72* (6), 455-482.

2. Ohtsubo, K.; Marth, J. D., Glycosylation in cellular mechanisms of health and disease. *Cell* **2006,** *126* (5), 855-867.

3. Rabenstein, D. L., Heparin and heparan sulfate: structure and function. *Natural Product Reports* **2002,** *19* (3), 312-331.

4. Xie, B.; Costello, C. E., Carbohydrate Structure Determination by Mass Spectrometry. *Carbohydrate Chemistry, Biology and Medical Applications* **2008**, 29-57.

5. Zhao, Y. J.; Singh, A.; Li, L. Y.; Linhardt, R. J.; Xu, Y. M.; Liu, J.; Woods, R. J.; Amster, I. J., Investigating changes in the gas-phase conformation of Antithrombin III upon binding of Arixtra using traveling wave ion mobility spectrometry (TWIMS). *Analyst* **2015,** *14* (20), 6980-6989.

6. Zhao, Y. J.; Singh, A.; Xu, Y. M.; Zong, C. L.; Zhang, F. M.; Boons, G. J.; Liu, J.; Linhardt, R. J.; Woods, R. J.; Amster, I. J., Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (1), 96-109.

7. Ly, M.; Leach, F. E., III; Laremore, T. N.; Toida, T.; Amster, I. J.; Linhardt, R. J., The proteoglycan bikunin has a defined sequence. *Nature Chemical Biology* **2011,** *7* (11), 827-833.

8. Yu, Y. L.; Duan, J. N.; Leach, F. E.; Toida, T.; Higashi, K.; Zhang, H.; Zhang, F. M.; Amster, I. J.; Linhardt, R. J., Sequencing the Dermatan Sulfate Chain of Decorin. *J. Am. Chem. Soc.* **2017,** *139* (46), 16986-16995.

9. Chi, L. L.; Amster, J.; Linhardt, R. J., Mass spectrometry for the analysis of highly charged sulfated carbohydrates. *Current Analytical Chemistry* **2005,** *1* (3), 223-240.

10. Chi, L. L.; Wolff, J. J.; Laremore, T. N.; Restaino, O. F.; Xie, J.; Schiraldi, C.; Toida, T.; Amster, I. J.; Linhardt, R. J., Structural analysis of bikunin glycosaminoglycan. *J. Am. Chem. Soc.* **2008,** *130* (8), 2617-2625.

11. Kailemia, M. J.; Li, L. Y.; Ly, M.; Linhardt, R. J.; Amster, I. J., Complete Mass Spectral Characterization of a Synthetic Ultralow-Molecular-Weight Heparin Using Collision-Induced Dissociation. *Analytical Chemistry* **2012,** *84* (13), 5475-5478.

12. Kailemia, M. J.; Patel, A. B.; Johnson, D. T.; Li, L. Y.; Linhardt, R. J.; Amster, I. J., Differentiating chondroitin sulfate glycosaminoglycans using collision-induced dissociation; uronic acid cross-ring diagnostic fragments in a single stage of tandem mass spectrometry. *European Journal of Mass Spectrometry* **2015,** *21* (3), 275-285.

13. Leach, F. E.; Riley, N. M.; Westphall, M. S.; Coon, J. J.; Amster, I. J., Negative Electron Transfer Dissociation Sequencing of Increasingly Sulfated Glycosaminoglycan Oligosaccharides on an Orbitrap Mass Spectrometer. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (9), 1844-1854.

14. Bin Oh, H.; Leach, F. E.; Arungundram, S.; Al-Mafraji, K.; Venot, A.; Boons, G. J.; Amster, I. J., Multivariate Analysis of Electron Detachment Dissociation and Infrared Multiphoton Dissociation Mass Spectra of Heparan Sulfate Tetrasaccharides Differing Only in Hexuronic acid Stereochemistry. *Journal of the American Society for Mass Spectrometry* **2011,** *22* (3), 582-590.

15. Wolff, J. J.; Amster, I. J.; Chi, L.; Linhardt, R. J., Electron detachment dissociation of glycosaminoglycan tetrasaccharides. *Journal of the American Society for Mass Spectrometry* **2007,** *18* (2), 234-244.

16. Wolff, J. J.; Laremore, T. N.; Busch, A. M.; Linhardt, R. J.; Amster, I. J., Electron detachment dissociation of dermatan sulfate oligosaccharides. *Journal of the American Society for Mass Spectrometry* **2008,** *19* (2), 294-304.

17. Wolff, J. J.; Laremore, T. N.; Busch, A. M.; Linhardt, R. J.; Amster, I. J., Influence of charge state and sodium cationization on the electron detachment dissociation and infrared multiphoton dissociation of glycosaminoglycan oligosaccharides. *Journal of the American Society for Mass Spectrometry* **2008,** *19* (6), 790-798.

18. Wolff, J. J.; Laremore, T. N.; Leach, F. E.; Linhardt, R. J.; Amster, I. J., Electron capture dissociation, electron detachment dissociation and infrared multiphoton dissociation of sucrose octasulfate. *European Journal of Mass Spectrometry* **2009,** *15* (2), 275-281.

19. Wolff, J. J.; Leach, F. E.; Laremore, T. N.; Kaplan, D. A.; Easterling, M. L.; Linhardt, R. J.; Amster, I. J., Negative Electron Transfer Dissociation of Glycosaminoglycans. *Analytical Chemistry* **2010,** *82* (9), 3460-3466.

20. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M., The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biological Chemistry* **2012,** *393* (11), 1357-1362.

21. Agyekum, I.; Patel, A. B.; Zong, C. L.; Boons, G. J.; Amster, I. J., Assignment of hexuronic acid stereochemistry in synthetic heparan sulfate tetrasaccharides with 2-O-sulfo uronic acids using electron detachment dissociation. *Int. J. Mass Spectrom.* **2015,** *390*, 163-169.

22. Leach, F. E.; Ly, M.; Laremore, T. N.; Wolff, J. J.; Perlow, J.; Linhardt, R. J.; Amster, I. J., Hexuronic Acid Stereochemistry Determination in Chondroitin Sulfate Glycosaminoglycan Oligosaccharides by Electron Detachment Dissociation. *Journal of the American Society for Mass Spectrometry* **2012,** *23* (9), 1488-1497.

23. Wolff, J. J.; Chi, L. L.; Linhardt, R. J.; Amster, I. J., Distinguishing glucuronic from iduronic acid in glycosaminoglycan tetrasaccharides by using electron detachment dissociation. *Analytical Chemistry* **2007,** *79* (5), 2015-2022.

24. Zaia, J.; Li, X. Q.; Chan, S. Y.; Costello, C. E., Tandem mass spectrometric strategies for determination of sulfation positions and uronic acid epimerization in chondroitin sulfate oligosaccharides. *Journal of the American Society for Mass Spectrometry* **2003,** *14* (11), 1270-1281.

25. Karlin, S.; Altschul, S. F., METHODS FOR ASSESSING THE STATISTICAL SIGNIFICANCE OF MOLECULAR SEQUENCE FEATURES BY USING GENERAL SCORING SCHEMES. *Proc. Natl. Acad. Sci. U. S. A.* **1990,** *87* (6), 2264-2268.

26. Karlin, S.; Altschul, S. F., APPLICATIONS AND STATISTICS FOR MULTIPLE HIGH-SCORING SEGMENTS IN MOLECULAR SEQUENCES. *Proc. Natl. Acad. Sci. U. S. A.* **1993,** *90* (12), 5873-5877.

27. Mackey, A. J.; Haystead, T. A. J.; Pearson, W. R., Getting more from less - Algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics* **2002,** *1* (2), 139-147.

28. Fenyo, D.; Beavis, R. C., A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry* **2003,** *75* (4), 768-774.

29. Field, H. I.; Fenyo, D.; Beavis, R. C., RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2002,** *2* (1), 36-47.

CHAPTER 4

SEQUENCING THE DERMATAN SULFATE CHAIN OF DECORIN**

Yu, Y. L*.; Duan, J. N.*; Leach, F. E.; Toida, T.; Higashi, K.; Zhang, H.; Zhang, F. M.; Amster, I. J.; Linhardt, R. J. *J. Am. Chem. Soc.* **2017,** *139* (46), 16986-16995.

* Jiana Duan and Yanlei Yu contributed equally to this work.

4.1 ABSTRACT

Glycomics represents one of the last frontiers and most challenging in omic analysis. Glycosylation occurs in the endoplasmic reticulum and the Golgi organelle and its control is neither well understood nor predictable based on proteomic or genomic analysis. One of the most structurally complex classes of glycoconjugates is the proteoglycans (PGs) and their glycosaminoglycan (GAG) side chains. Previously, our laboratory solved the structure of the chondroitin sulfate chain of the bikunin PG. The current study examines the much more complex structure of the dermatan sulfate GAG chain of decorin PG. By utilizing sophisticated separation methods followed by compositional analysis, domain mapping and tandem mass spectrometry coupled with analysis by a modified genetic algorithm approach, the structural motif for the decorin dermatan sulfate chain was determined. This represents the second example of a GAG with a prominent structural motif, suggesting that the structural variability of this class of glycoconjugates is somewhat simpler than had been expected.

4.2 INTRODUCTION

The recent and rapid progress in genomics and proteomics has not been matched in glycomics, corresponding to the largest portion of the metabolome.[1] One reason for the slow progress in establishing glycomes of various organisms is the complexities associated in the structural characterization and sequencing of complex carbohydrates.[2] Genomic sequencing relies on amplification methods that can also be applied to the proteomic sequencing as the result of the one-gene to one protein paradigm.[3] Unfortunately, posttranslational modification of the proteome in the case of glycosylation involves non-template driven biosynthesis beginning in the endoplasmic reticulum (ER), continuing in the Golgi and completed in catabolic remodeling outside the cell.[4] Thus, structural characterization and sequencing need to be carried out using conventional analytical chemistry relying primarily on separation and spectroscopy methods.

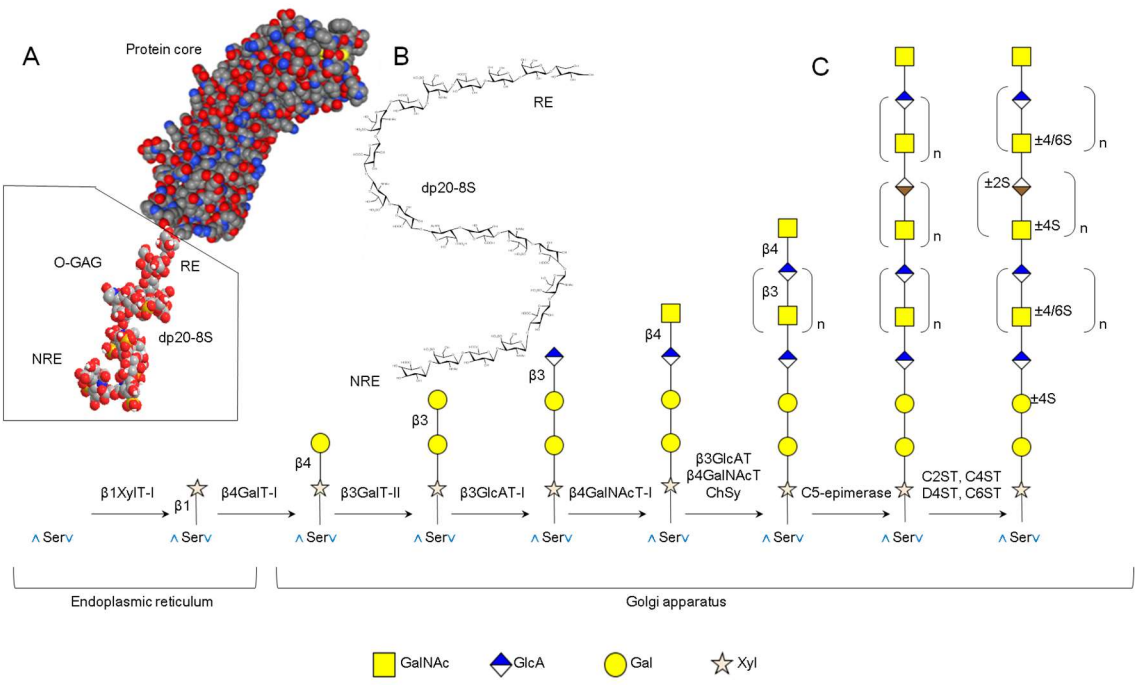


Figure 4.1. Modeled structure and biosynthetic pathway of decorin glycosaminoglycan. **A**. Space filling structure of decorin PG. Decorin core protein from PDB (1XCD.pdb). Carbons (gray), hydrogens (white), oxygens (red), nitrogens (blue), and sulfurs (yellow) are shown. *O*-linked GAG chain (dp 20-8S) are shown with the reducing end (RE) and non-reducing end (NRE). **B**. Chemical structure of GAG chain dp 20-8S with a tetrasaccharide linkage region (GlcA-Gal-Gal-Xyl) at the RE. **C**. Biosynthetic pathway for chondroitin sulfate/dermatan sulfate GAG. The GAG on a serine residue of the core protein, is synthesized in a pathway that begins in the endoplasmic reticulum and concludes in the Golgi apparatus. The biosynthetic enzymes are: β1XylT-I, β-xylosyl transferase I; β4GalT-I, β-4-galactosyl transferase I; β3GalT-II, β-3-galactosyl transferase II, β3GlcAT-I, β-3-glucuronosyl transferase I; β4GalNAcT-I, β-4-*N*-acetyl galactosaminyl transferase I; β3GlcAT, β-3-glucuronosyl transferase; β4GalNAcT, β-4-*N*-acetyl galactosaminyl transferase; ChSy, chondroitin synthases; C5-epimerase; C2ST, 2-*O*-sulfotransferases, C4ST, chondroitin 4-*O*- sulfotransferases, D4ST, dermatan 4-*O*-sulfotransferases, C6ST, 6-*O*-sulfotransferases.

Proteoglycans (PGs) are among the most structurally complex glycoconjugates and are polydisperse, microheterogeneous mixtures having average molecular mass ranging from 25 to 2,500 kDa.[5,6] These PG glycoconjugates are biosynthesized in three steps (Figure 4.1). The first is the template driven synthesis of the core protein in the rough ER, the second is the installation of tetrasaccharide linkage regions, on specific serine residues of the core protein, and the third is transit through the Golgi and extension of the glycosaminoglycan (GAG) polysaccharide chains and the structural modification of their saccharide residues through epimerization and sulfation (Figure 4.S1).[7-9] The structural complexity of PGs are associated with: (1) the occupancy of GAGylation sites; (2) the type of GAG chains, *i.e.,* chondroitin sulfate, dermatan sulfate, heparan sulfate, *etc.*; (3) the length of each GAG chain, *i.e.,* degree of polymerization (dp) or number of saccharide residues; and (4) the structure or sequence of each individual GAG chain. The simplest 25 kDa PG, bikunin, has a single GAGylation site occupied by a ~6 kDa chondroitin/chondroitin-4-sulfate GAG chain, of dp27-39, with a single well defined sequence motif.[10] In contrast, one of the more complex PGs, aggrecan has a molecular weight of 2,500 kDa, has up to 160 GAGylation sites occupied by either ~100 chondroitin/chondroitin-4 and/or chondroitin-6-sulfate GAG chains of ~80 dp, and ~60 keratan/keratan sulfate (6-*O*-sulfo-galactose and/or 6-*O*-sulfo-*N*-acetylglucosamine) GAG chains of ~ dp40,[11] with still unknown sequence motifs.

Ongoing research in our laboratory has focused on the study of the structural glycomics of PGs. We began with the simplest PG, bikunin[12] a serine protease inhibitor with an important role in inflammation,[13] potentially having ~10^{11} GAG sequences, and demonstrated that it had a singular sequence motif (Figure 4.S2).[10] Over the past decade

our laboratory[14,15] and others[16-19] have laid the biochemical groundwork to examine the structure and sequence of the next simplest PG, decorin. These studies included disaccharide compositional analysis, linkage region variability analysis and domain mapping. The decorin GAG chain structure elucidated in these studies represents a composite average and does not necessarily correspond to an actual sequence nor does it provide any information of sequence variability (*i.e.,* the number of potential sequences that are actually present). Decorin with its single GAGylation site, occupied by a 4-*O* and/or 2-*O* and/or 6-*O*-sulfo dermatan / chondroitin dp14-40 GAG chain (Figure 4.S3), potentially having ~10^{18} GAG sequences (Table 1). This large number of sequence permutations is due to both the 8 different disaccharide structures, comprised of 1→3-linked L-iduronic acid (IdoA) (with and without 2-*O*-sulfo (2S) groups) and D-glucuronic acid (GlcA) and 1→4-linked *N*-acetyl-D-galactosamine (GalNAc) (with/without 4S and/or 6S) and with decorin's requisite polydispersity.[14,15] Thus, we anticipated that our efforts for the direct sequencing of the decorin GAG chain would be roughly a million times more difficult than our first successful sequencing of bikunin.[10]

DP	Disaccharides	Total Possible	Limited Number of SO3	Disaccharide Analysis Restrictions	Tandem MS Restriction
14	7	2.10E+06	2.77E+04	4.12E+02	2
15	7	8.39E+06	1.43E+05	1.85E+03	26
16	8	1.68E+07	1.88E+05	2.11E+03	27
17	8	6.71E+07	9.54E+05	4.46E+03	29
18	9	1.34E+08	1.26E+06	5.04E+03	30
19	9	5.37E+08	6.33E+06	2.16E+04	61

20	10	1.07E+09	8.48E+06	2.50E+04	63
21	10	4.29E+09	4.19E+07	5.33E+04	67
22	11	8.59E+09	5.67E+07	6.12E+04	69
23	11	3.44E+10	2.78E+08	2.54E+05	6.38E+02
24	12	6.87E+10	3.79E+08	2.96E+05	6.74E+02
25	12	2.75E+11	1.84E+09	6.35E+05	7.49E+02
26	13	5.50E+11	2.52E+09	7.37E+05	7.88E+02
27	13	2.20E+12	1.22E+10	2.98E+06	1.61E+03
28	14	4.40E+12	1.68E+10	3.51E+06	1.69E+03
29	14	1.76E+13	8.07E+10	7.56E+06	1.86E+03
30	15	3.52E+13	1.12E+11	8.84E+06	1.95E+03
31	15	1.41E+14	5.35E+11	3.51E+07	1.64E+04
32	16	2.81E+14	7.45E+11	4.16E+07	1.75E+04
33	16	1.13E+15	3.54E+12	8.98E+07	1.98E+04
34	17	2.25E+15	4.96E+12	1.06E+08	2.10E+04
35	17	9.01E+15	2.35E+13	4.14E+08	4.32E+04
36	18	1.80E+16	3.30E+13	4.93E+08	4.58E+04
37	18	7.21E+16	1.56E+14	1.07E+09	5.13E+04
38	19	1.44E+17	2.20E+14	1.26E+09	5.42E+04
39	19	5.76E+17	1.03E+15	4.90E+09	4.29E+05
40	20	1.15E+18	1.46E+15	5.85E+09	4.60E+05
	Totals	**1.98E+18**	**2.94E+15**	**1.43E+10**	**1.17E+06**

Table 4.1. Showing the number of possible permutations for a decorin glycan of a specific chain length with various levels of restrictions set. Six columns headings are

shown. 1. Degree of polymerization (dp); 2. Number of disaccharides of specified dp; 3. Number of possible permutations for SO_3 modifications for specified dp assuming all possible dermatan sulfate modifications (0S, 2S, 4S, 6S, 2S4S, 2S6S, 4S6S, 2S4S, 2S4S6S). Calculations for total possible number of permutations was calculated as summation of a binomial combination using n-choose-k [n!/k!(n-k)!], where n is the total number of saccharide units and k is the number of modifications with k ranging from 0 to 3 times the number of disaccharides; 4. Limited number of permutations based on composition assignments done with MS1 (FT-ICR-MS and Orbitrap-MS) at high mass accuracy, a limited number of SO_3 modifications are possible with respect to chain length. Number of possible SO_3 is reduced to number of dissacharides/2 – 4 to number of dissacharides/2 – 2. N-choose-k calculations are still used but with a significantly reduced range in k; 5. Disaccharide analysis restrictions possible sulfo group modifications on any disaccharide to four possible combinations: 0S, 2S4S, 6S, 4S. Tandem MS analysis restrictions where MS^2 reveals the prevalence of a characteristic, multi-charged peak that suggest a single sulfate modification per disaccharide from the non-reducing end that extends 3 to 10 disaccharide units (for dp14 to dp40).

Decorin is the simplest pericellular PG member belonging to the small leucine-rich proteoglycan (SLRP) family. It was named based upon its property of "decorating" collagen fibrils in the skin and tendons and control of fibrillogenesis.[20-23] Decorin has been called the guardian of the extracellular matrix (ECM) resulting from its role as a pan-inhibitor of tyrosine kinase signaling and in this role decorin displays a large ECM interactome of importance in controlling tumor growth, angiogenesis and autophagy.[24] While many of decorin's biological activities are associated with its banana-shaped core protein, others, particularly its ability to modulate matrix maturation and its interaction with ECM enzymes, such as metalloproteinases and growth factors, require its GAG chain.[23] The GAG chains of endothelial decorin are also important participants in human

infection by the tick born spirochete, *Borrelia bergdorferi* that causes Lyme disease, by serving as a receptor for its surface proteins facilitating tissue colonization.[25,26] The flexible dermatan sulfate chain of decorin PG,[27] particularly the highly sulfated domains rich in iduronic acid, are important features for protein interaction associated with biological activity.[28]

The current study examines porcine skin decorin, available in multigram quantities, and undertakes its extensive fractionation to prepare a less heterogeneous mixture of decorin chains with an average level of sulfation but with a relatively small chain length (low dp), as determined by analytical polyacrylamide gel electrophoresis (PAGE). The disaccharide composition, linkage region structures, and domain structures of these chains were determined by mass spectrometry. The accurate mass measurement of GAG chains by Fourier transform mass spectrometry (FTMS) enabled the determination of polymerization and sulfation extent. Collisional dissociation tandem mass spectrometry was used to determine the pattern of sulfo groups through the decorin GAG chain and afforded a prominent sequence motif.

4.3 RESULTS AND DISCUSSION

Linkage Region. The decorin GAG chain is biosynthesized on serine 34 near the N-terminus of its core protein (Figure 4.1A). The linkage region tetrasaccharide of porcine skin decorin, →4) GlcA (1→3) galactose (Gal) (±4S) (1→3) Gal (1→4) xylose (Xyl) (±2 phospho (P)) (1→, assembled in the ER is variable with ~30% of the chains contain a 4-S-Gal and ~5% of the chains contain a 2-P-Xyl, (Figure 4.S4 & S5).[15] Chain extension on this linkage region tetrasaccharide occurs in the Golgi compartment resulting in a linear

polysaccharide chain having 14-40 saccharide residues (Figure 4.1B). C5-epimerization, converting GlcA to IdoA, and the introduction of sulfo groups to the 4-, 6- and 2- positions afford the mature decorin dermatan sulfate GAG chain (Figure 4.1C). Not all modification steps are complete and results in considerable structural hetereogeneity. Moreover, enzyme selectivity and/or other unknown control factor(s) results in the formation of structural domains within the mature decorin dermatan sulfate GAG chains. The GAG chains are typically released from the core protein through β-elimination under reducing conditions so that each chain carries a xylitol (Xyt) at its reducing end (Figure 4.1B). The released GAG chains present such a complex mixture of polymer lengths, sulfation levels, and domains that it is not possible to apply current mass spectrometry methods to determine their sequence. A fractionation approach was designed to prepare a representative set of chains of relatively short chain length but with average sulfation density for sequence analysis (Figure 4.2). Ten kg of porcine skin (wet weight) afforded 2 g of decorin PG, 1 g of decorin GAG chains, and ultimately 1 □g of size-uniform and charge-uniform GAG chains for sequencing.

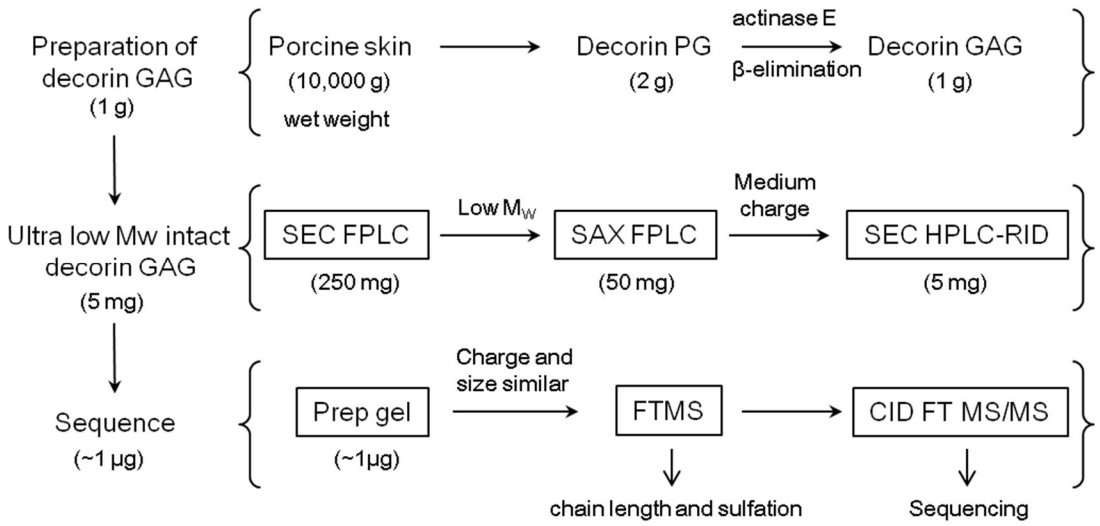


Figure 4.2 Decorin flow chart for solving structure. From 10,000 g of wet porcine skin 2 g of decorin proteoglycans of high purity was obtained. Proteolysis with actinase E afforded decorin peptidoglycan which was converted to 1 g of decorin GAGs by reductive β-elimination. Size exclusion chromatography was applied to obtain low molecular weight decorin and strong anion exchange was applied to get medium charge fractions. SEC with refractive index detector on HPLC was applied to online separate decorin fractions. Continued elution preparative PAGE was next done to get size and charge similar fractions for FT and FT-MS/MS analysis.

Composition and Domain Mapping. The initial mixture of decorin GAG chains released from decorin PG had an average molecular mass of 22 kDa, a chain length ranging from dp14-dp290 based on the lowest and highest molecular weight acquired from GPC, and an average sulfation density of 0.98 sulfo groups/disaccharide repeating unit (Figure 4.3). Disaccharide compositional analysis, performed by chondroitin lyase catalyzed depolymerization of the decorin GAG to unsaturated disaccharides (with a non-reducing terminal □UA, deoxy-□-L-threo-hex-4-enopyranosiduronic acid) and HPLC-MS analysis, afforded a composition of 96.8 mol % □UA(1→3)GalNAc4S, 2.0 mol %

□UA2S(1→3)GalNAc4S, 1.0 mol % □UA(1→3)GalNAc6S and 0.2 mol % □UA(1→3)GalNAc. Next, we examined the AC/B domain structure of the decorin GAG chain using NMR spectroscopy, which showed an IdoA/IdoA2S: GlcA ratio of 3:1 (Figure 4.3). The B domain contains IdoA or IdoA2S residues and is susceptible to treatment with endolytic chondroitin B lyase, while the AC domain is susceptible to treatment with endolytic chondroitin AC lyase (Figure 4.S6-S8).[29-31] Thus, exhaustive treatment of decorin GAG with chondroitin B lyase and chondroitin AC lyase and recovery of intact chains of reduced size affords AC and B domains, respectively. Disaccharide analysis shows that the AC domains exclusively contain □UA(1→3)GalNAc4S, □UA(1→3)GalNAc6S and □UA(1→3)GalNAc at 93.2 mol %, 6.0 mol % and 0.8 mol %, respectively, and the B domain exclusively contains only □UA(1→3)GalNAc4S at 95.3 mol %, and □UA2S(1→3)GalNAc4S at 4.7 mol % (Figure 4.S8)□□□ This analysis is consistent with the presence of 2-*O*-sulfo groups present only on IdoA residues and only in B domains. As expected, the B domains were often >dp10 and were on average ~3-times longer than the AC domains which were most frequently dp6 (Figure 4.S7). Linkage region analysis confirmed the presence of two major reduced structures from the linkage region domain, →4) GlcA (1→3) Gal (±4S) (1→3) Gal (1→4) Xyt comprising >90 mol % of the decorin chains (Figure 4.3).

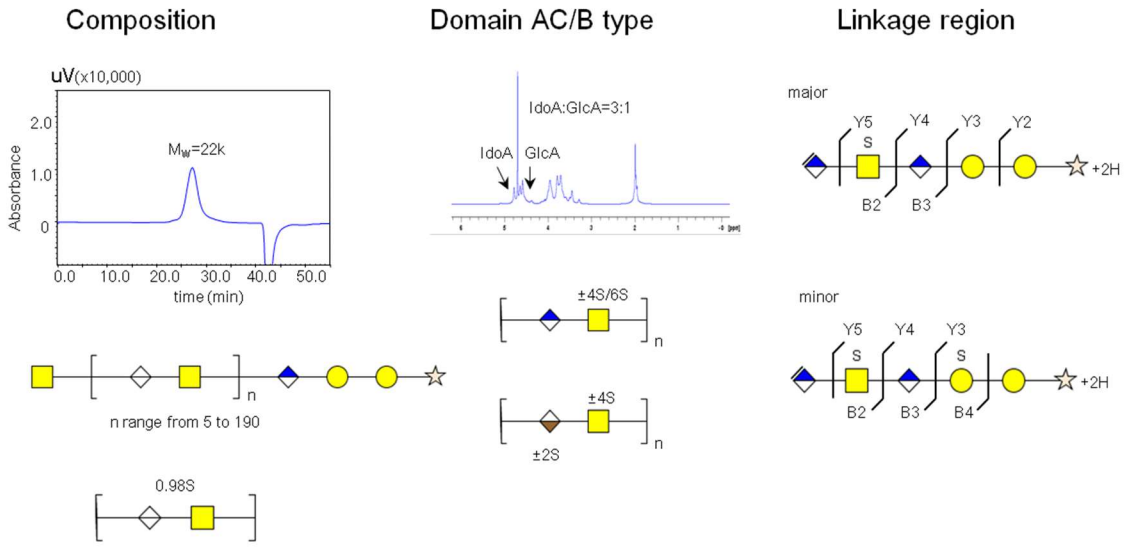
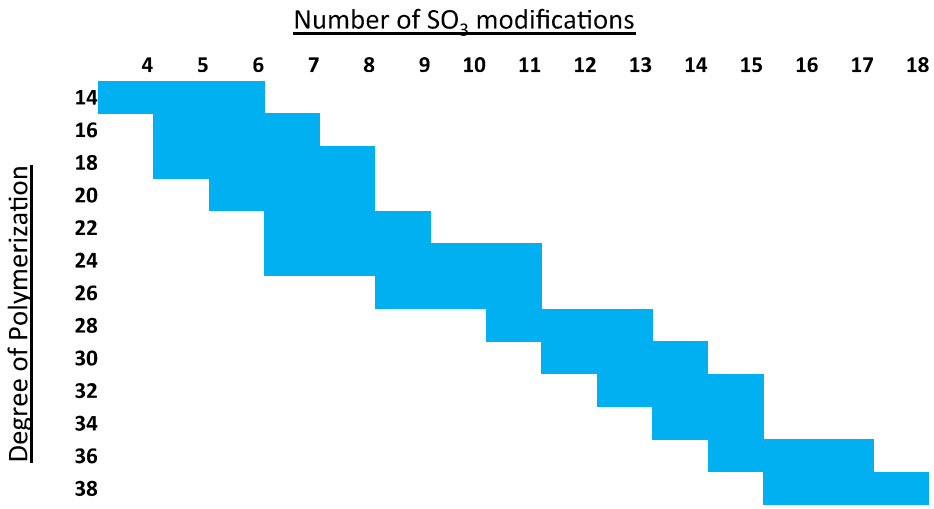


Figure 4.3. Decorin GAG compositional, domain mapping and linkage analysis. Decorin GAG molecular weight was determined by GPC-HPLC. Disaccharides analysis showed average sulfation of per disaccharide was 0.98, the chain length was ranged from dp14-dp290 and average was dp92. The ration of the IdoA-H1 and GlcA-H1 was calculated as 3:1 based on ^{1}H-NMR. Disaccharides analysis of AC-type and B-type domain revealed that 2S4S only exists on B-type domain, so did 6S and 0S only exist on AC-type domain. Completed digestion of decorin GAG afforded linkage region, major linkage region GlcA-GalNAc4S-GlcA-Gal-Gal-Xyt and minor linkage region GlcA-GalNAc4S-GlcA-Gal4S-Gal-Xyt.

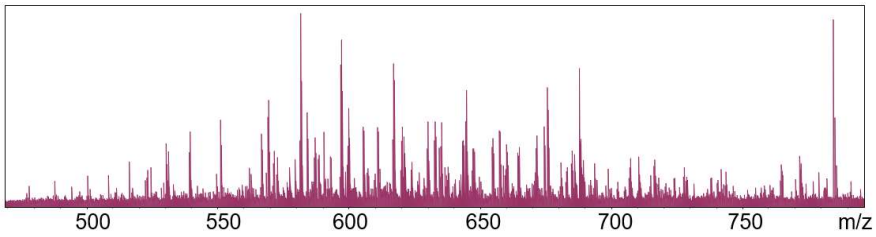
Simplifying Mixture Complexity by Fractionation. Our experience in MS sequencing the much less complex bikunin GAG chain suggested that while a sulfation level of 0.98 *O*-sulfo groups/disaccharide repeating unit of the decorin GAG was a tractable problem, the large chain length up to dp290 (Figure 4.S3) would challenge the limits of current MS capabilities. Moreover, the presence of more than 10-20 prominent molecular ions in a fraction make the selection of one having the appropriate intensity for MS/MS analysis problematic. We set out to fractionate the decorin GAG chains to obtain relatively

homogenous fractions of chains of average sulfation level but of sizes of dp<36 to generate a sample set suitable for mass spectrometry. Decorin GAG (1 g) of Mw (avg) 22 kDa (Figure 4.3) was first fractionated by size exclusion chromatography (SEC) fast performance liquid chromatography (FPLC) and approximately the last third of the eluting sample, corresponding to Mw (avg) 19 kDa was collected (Figure 4.2). Next, 250 mg of this fraction was applied to strong anion exchange (SAX)-FPLC and the center third of the peak was collected and subjected to SEC-HPLC to again enrich 5 mg of the small chains (Figure 4.S9). Finally, preparative PAGE was applied to obtain 135 fractions that were analyzed by analytical PAGE (Figure 4.S10) of microgram quantities of fractions ranging in estimated size from 3 kDa to 32 kDa (Table S1). Orbitrap FTMS analysis was then undertaken on selected PAGE fractions ranging from #38 (Mw~4.2 kDa) to #60 (Mw~8.7 kDa) to determine the intact mass of fraction components and to assess the suitability of these and neighboring fractions for MS/MS sequencing (Figures S11-S18 and Table S2-S9). Accurate masses were obtained for 57 molecular ions ranging from dp20 to dp44 and carrying 8-20 sulfo groups with only a small amount of sodium adduction observed. The complexity of these fractions with 10-20 molecular ions detected in each and sufficient S/N suggested that these fractions would be suitable for sequencing by MS/MS.

A



B



Composition	M/Z	Z	ME	Composition	M/Z	Z	ME
dp10-3S	675.4622	3	-1.28	dp20-7S	706.4574	6	-1.83
dp11-4S	769.8082	3	-0.38	dp20-7S	605.3913	7	-1.41
dp12-3S	801.8335	3	-0.06	dp20-8S	616.8147	7	-1.42
dp12-4S	621.1122	4	-0.64	dp20-8S (1 Na)	619.9536	7	-1.91
dp13-4S	671.8813	4	-1.74	dp22-6S	659.4364	7	1.05
dp13-5S	553.2949	5	-1.80	dp22-8S	670.972	7	-1.93
dp14-4S	715.8899	4	-0.84	dp22-8S	586.975	8	-1.23
dp14-5S	588.5015	5	-1.31	dp22-8S (1Na)	674.1138	7	0.28
dp16-5S	664.3242	5	-0.59	dp22-9S	596.9692	8	-0.81
dp16-6S	566.7611	6	-1.64	dp22-9S (1Na)	599.7176	8	-0.76
dp16-6S	680.3156	5	-0.54	dp24-7S	713.7093	7	-0.57
dp18-5S	616.6203	6	-1.41	dp24-9S	572.651	9	-1.59
dp18-5S	740.1459	5	-1.31	dp24-9S	644.359	8	-0.38
dp18-6S	629.9466	6	-1.18	dp24-10S	581.5353	9	-1.63
dp18-7S	551.2323	7	-1.86	dp24-10S	654.3536	8	-1.58
dp18-7S (1Na)	646.9367	6	-0.68	dp26-9S	614.7744	9	-1.58

dp18-8S	562.6549	7	-1.49	dp26-10S	623.6583	9	-1.94
dp20-6S	693.1323	6	-0.36	dp26-10S	561.1931	10	0.59

Figure 4.4. Complexity of decorin GAG chain mixtures over a range of chain sizes. A. Accurate mass measurement in the MS1 using FT-ICR MS and Orbitrap-MS makes composition assignment possible. Compositions calculated from aggregated MS1 data show a limit in the overall number of SO_3 modifications that can exist for a specific dp. **B.** Example of composition assignments for decorin fraction 39 using FT-ICR MS. All compositions are within 2 ppm mass error.

Sequencing. FT-ICR MS was applied to neighboring PAGE fraction #39. Initial MS analysis provided molecular compositions which were then selected for sequencing by collisional dissociation. The dp range observed in fraction #39 (dp14-26) by FT-ICR MS (Figure 4.4A) was slightly wider than had been observed in fraction #38 (dp16-24) by Orbitrap FTMS (Figure 4.S11, Table S2). At the MS level, molecular ions corresponding to 25 unique compositions were detected and four were found suitable for further MS/MS analysis (Figure 4.4B, Figure 4.S19-S20, Table S10-S11). Ions were not selected for MS/MS analysis for several reasons including: a) lack of clean quadrupole mass selection where more than one composition was isolated; b) low intensity precursor ions that did not provide fragment ions above the limit of detection; or c) depletion of the same fraction from multiple MS experiments. A sample annotated spectrum of decorin MS/MS is shown in Figure 4.5 where CID-FT-ICR-MS/MS of the molecular ion m/z 616.8147, corresponding to dp20-$7S^{7-}$, resulted in a spectrum rich in glycosidic bond cleavages suitable for sequence determination. Additional sequence analysis of other tandem MS from fraction #39 are available in supplemental information. Complete composition and

tandem MS analysis of fractions #35 and #51 are also provided (Figures S21-S32, Tables S12-S23).

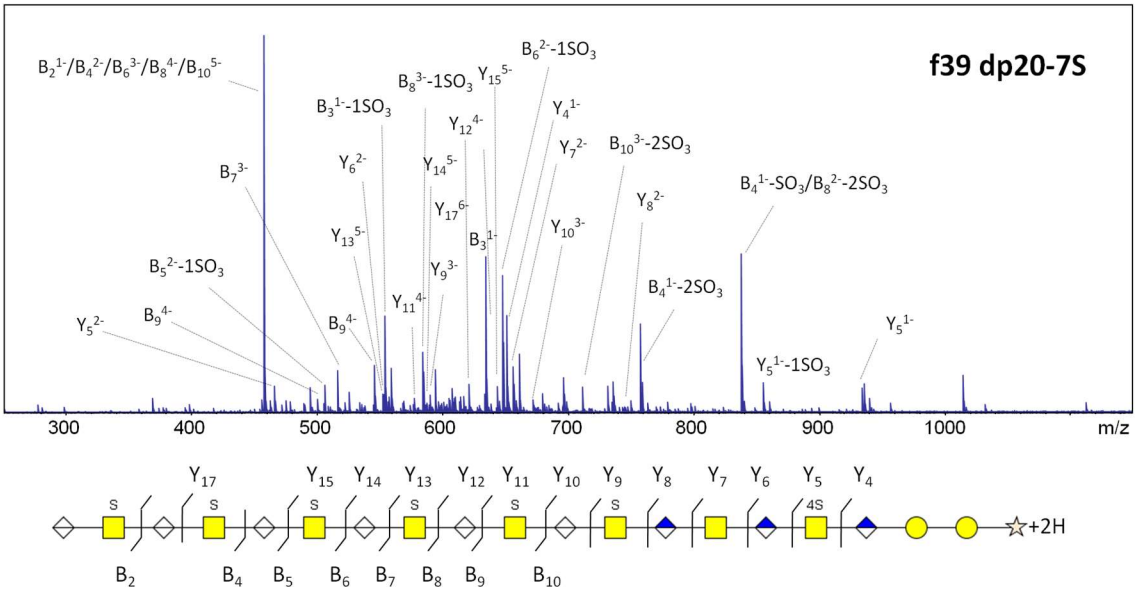


Figure 4.5. Tandem MS analysis of decorin GAG chain DP20-7S^{7-}. Structure was determined using in-house GAG algorithm software and corresponding MS^2 cleavages are shown.

Matching adjacent glycosidic fragments ($B_n + B_{n+1}$, $Y_n + Y_{n+1}$, etc.,) would differ by 176.0321 Da for hexuronic acid and 203.0793 Da for *N*-acetyl galactosamine as observed previously in bikunin glycan analysis.[10] Non-matching adjacent glycosidic fragments (ex. $B_n + C_{n+1}$ or $C_n + B_{n+1}$, $Y_n + Z_{n+1}$, etc.) differ by an additional ±18.0106 Da. Sulfo group modifications were assigned to sugar residues based on the addition of 79.9568 Da between adjacent sets of glycosidic fragments. Based on disaccharide analysis, sulfo modifications exist as only 2-*O*- on hexuronic acids but as 4-*O*- or 6-*O*- sulfo groups on *N*-acetyl galactosamine. A 4-*O*- modification was determined for the reducing end linker region

(GlcA-GalNAc4S-GlcA-Gal-Gal-Xyt) but additional differentiation of 4-*O*- and 6-*O*- sulfo groups depended on the presence of diagnostic cross-ring fragmentations. CID ion activation was capable of breaking single bonds between residues, providing glycosidic fragments but lacked sufficient energy to break multiple bonds across the hexose sugar residue, yielding no diagnostic cross-ring fragments to localize sulfo moeties, except in rare cases. One rare case is shown in Figure 4.S23 and Table S14 where cross-ring fragments are used to differentiate 4-*O*- and 6-*O*- sulfo modifications. Presence of the $^{1,4}X_6$ fragment ion with two sulfo groups sandwiched between the Y_6 ion containing one sulfo-group and Y_7 ion containing two sulfo groups is evidence of a 6-*O*- modification. Regions where cross ring fragments were identified (both diagnostic and non-diagnostic) are provided in the supplemental section.

Assigned decorin structures show a consistent run of sequential glycosidic fragments from both the reducing and non-reducing end. Assigned glycosidic fragments with the highest signal intensity and greatest occurrence were B and Y ions. Low abundance C fragments (<5%) were observed in nearly all spectra as well while Z fragment ions were rarely assigned. The intense peak at m/z 458.06 observed in every tandem MS of decorin regardless of fraction or composition matches the mass of a B-fragment ion for a singly-sulfated disaccharide residue, or a polymer chain of the same residue, $(HexA\text{-}GalNAc\text{-}S - H)_n^{n-}$. Examination of the isotopic distribution shows the possibility of multiple charge states from -1 to -11 in some spectra, suggesting a repetitive 1-sulfo group per disaccharide chain for up to 22 sugar residues. Comparison of this isotope distribution between MS/MS spectra shows the 1- charge state as dominant for shorter chains (dp18) whereas the 4- charge state is most abundant for the highest chain. Loss of the sulfo modification was

observed most commonly for glycosidic B-fragments (B_1-B_8) but typically appeared at intensities of 3-22% of the fully modified counterpart. The C-5 stereochemistry of acidic residues, (IdoA vs. GlcA) could not be determined with our current tandem mass spectrometry method. Work from Zaia et al.[32, 33] suggests a correlation between C-5 epimerization and abundance of $^{0,2}X_n$ and Y_n ions for CS/DS but lacks definitive measurements for application in long chain glycans such as decorin. Differentiation of the C-5 uronic acid stereochemistry in CS/DS is possible using electron detachment dissociation (EDD), based on a diagnostic ratio of fragment ion relative intensity that was determined using principal component analysis (PCA).[33-35] Current work is focused on extending the diagnostic ratios approach to longer oligomers. While future efforts will be required to establish the definitive sequence for porcine skin decorin GAG chains, a major structural motif can be derived based on both biochemical analysis and MS sequencing (Figure 4.6).

Non-reducing end motif. Sequence analysis shows homogeneity in the non-reducing end of porcine decorin GAG. The HexA-GalNAc-S disaccharide motif is consistently repeated at the non-reducing end with no evidence of alternative structures. Although information to determine C-5 uronic sugar stereochemistry is lacking, intense series of B fragments and/or C fragments are present in all decorin tandem MS for the HexA-GalNAc-S repeating unit. The overall percentage of HexA-GalNAc-4S disaccharide as determined by disaccharide analysis increases with respect to chain length. These findings are reflected in the tandem MS results with the total number of observed sequential B-ions that favor the HexA-GalNAc-S pattern increasing when chains are longer.

Reducing end motif. Observable variations in sulfation patterns of different sequences occurs primarily at the reducing end. HexA-GalNAc6S, IdoA2S-GalNAc4S and GlcA-GalNAc disaccharide variants occur infrequently as suggested by disaccharide analysis and are minor components in the overall sequence. Sequences derived from tandem MS data show that these alternatives occur 0 to 2 times per chain. Sequential series of Y fragments suggest variability in the region closest to the reducing end after the GlcA-GalNAc4S-GlcA-Gal-Gal-Xyt linker region. An unmodified disaccharide unit exists 1-4 disaccharide units after the linker region. The IdoA2S-GalNAc4S modification exists in a similar region. Cross-ring fragments that validate the presence of 6-*O*-sulfo modification on the GalNAc (hence a GlcA-GalNAc6S unit) are only observed near the reducing end. Sequential Y-ions near the reducing end exist from 2-30% relative ion intensity but exhibit no common intensity-dependent pattern (unlike B-ions at the non-reducing end). The reducing end is the region where all variations to the HexA-GalNAc-S pattern exist but are observed in no specific order.

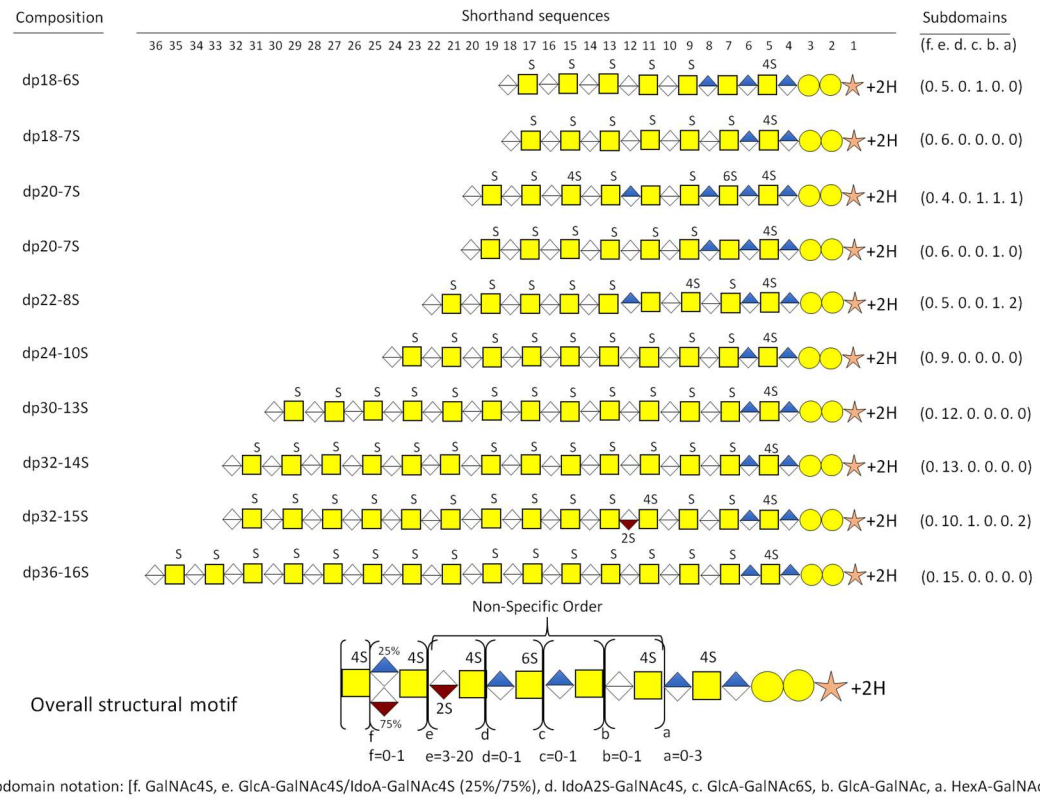


Figure 4.6. **Motif for decorin GAG chains from dp14-dp40.** Disaccharide analysis reveals a relatively small fraction of disaccharides having modifications that are not 4S on the GalNAc. MS^2 shows that the majority of these GalNAc4S exist near the non-reducing end. Structural variability occurs primarily near the linker region. GlcA2S-GalNAc4S and GlcA-GalNAc are both observed in the MS^2 but never more than one instance per chain. 6-*O*-sulfation of the GalNAc is more difficult to observe using the CID ion activation method as differentiation from the 4-*O*-sulfation requires specific cross-ring cleavages. Electron-based methods have proved to be more useful but have yet to be successful on highly heterogeneous samples such as decorin.

4.4 EXPERIMENTAL SECTION

Materials. All chemicals and reagents were obtained from commercial sources and used as received unless otherwise specified. Chondroitin sulfate lyase ABC from *Proteus vulgaris* was expressed in our laboratory using *Escherichia coli* strain provided by

Professor Miroslaw Cygler (College of Medicine, University of Saskatchewan) and chondroitin sulfate lyase II from *Arthrobacter aurescens* was expressed in *E. coli*.[36]

Preparation of decorin glycosaminoglycan. Decorin proteoglycan was purified from porcine skin.[15] Decorin PG fraction was proteolyzed by a 5% (w/w) actinase E digestion at pH 8.0 in 50 mM Tris-HCl in sodium acetate. The enzymatic reaction proceeded at 55 °C for 24 h and was then isolated from the digestion mixture by strong-anion exchange spin column. Spin columns were pre-equilibrated with 8 M urea containing 2% (w/v) CHAPS and centrifuged at 500 × *g* for 5 min. The bound pG was washed once with 8 M urea containing 2% (w/v) CHAPS and three times with 50 mM NaCl. Decorin pG was eluted with 2 M NaCl, desalted using a 3 kDa MWCO centrifugal filter and lyophilized. The GAG component of decorin was released by base-catalyzed β-elimination under reducing conditions. Samples were dissolved in 0.2 M NaOH solution containing 1% $NaBH_4$. The reaction was allowed to proceed overnight at 4 °C and neutralized with 1 M hydrochloric acid. The resulting GAG mixture was purified using a 3 kDa MWCO centrifugal filter.

Linkage region analysis of decorin glycosaminoglycan. Approximately 200 µg of decorin glycosaminoglycan was digested completely using 200 mU chondroitin sulfate lyase ABC at 37 °C for 18 h, digested sample was then purified by a 3 kDa MWCO spin column filter to isolate reducing ends and lyophilized for LC-MS analysis.

Isolation and preparation of low molecular weight and medium charge GAG fractions. Size exclusion chromatography (SEC) was performed on ÄKTApurifier (Fast protein liquid chromatography, FPLC, GE Healthcare Bio-Science) using pre-packed superdex S75 column with a sample injection volume of 200 µL and a flow rate of 0.5 mL/min. The

mobile phase consisted of 0.2 M ammonium bicarbonate. Fraction collector (Frac 920) was set to 2 min in conjunction-accumulated fractions. GAG concentration in each fraction was determined by micro-carbozole assay. Strong anion exchange chromatography (SAX) was performed on ÄKTApurifier using Hiprep Q HP 16/10 column with a sample injection volume of 25 mL and a flow rate of 3 mL/min. The mobile phase was sodium chloride and water, gradient wash from 0 M to 2 M sodium chloride in 15 column volumes was applied. Further SEC fractionation was performed on HPLC (Shimadazu) using Superdex increase 75 10/300 GL (GE Healthcare) with refractive index detector.

Fractionation of decorin glycosaminoglycan by continuous elution PAGE. A gel of 10-cm column height with 4 mL of 15% total acrylamide monomer resolving solution was allowed to polymerize overnight with 4 µL TEMED and 12 µL 10% (w/v) ammonium persulfate and was cast in a Mini Prep column with a 7 mm internal diameter (Bio-Rad). Above the polymerized resolving gel, 1 mL of 5% total acrylamide monomer stacking gel was cast. An aliquot of 1 mg purified decorin glycosaminoglycan was loaded in a solution of 10 µg/mL (w/v) phenol red and 25% (w/v) sucrose. Electrophoresis was performed for 8 h at a constant power of 1 W with a peristaltic pump (Econo pump, Bio-Rad) set to 0.08 mL/min and fraction collector (Model 2110, Bio-Rad) set to 3 min in conjunction-accumulated separating fractions from the Mini Prep cell (Bio-Rad). Buffer salts from electrophoresis for each fraction were removed by strong anion exchange column (High-Capacity Mini-Q, Satorius) and thoroughly desalted by LC-grade water washed with 3 kDa spin column (Millipore). The extent of separation was visualized by 15% total acrylamide monomer solution using native mini-slab PAGE stained with Alcian blue and molecular

weight of the fractionated GAG was estimated on PAGE densitometry against heparin ladder using UN-SCANIT (Silk Scientific).

Orbitrap FTMS analysis of decorin glycosaminoglycan. Glycosaminoglycans were analyzed in the negative-ion mode by electrospray ionization on a Thermo Scientific LTQ Orbitrap XL FT mass spectrometer with a standard, factory-installed ion source (Thermo Scientific). External calibration of mass spectra produced a mass accuracy of <3 ppm. Samples were dissolved in 50% aqueous methanol with 0.1% formic acid and were delivered by an Agilent 1200 nano-LC pump at a flow rate of 50 μL/min. Mass spectra were acquired at a resolution of 60000, detection range m/z 400-2000, and the charge deconvolution was performed manually with electronic spreadsheets. Acquisition parameters used to prevent in-source fragmentation included spray voltage -4.2 kV, capillary voltage -15 V, tube lens voltage -100 V, capillary temperature 250 °C, sheath flow rate 25, and auxiliary gas flow rate 5. For collision-induced dissociation (CID) MS/MS of the linkage region, parent ions were fragmented with a specified collision energy of 35 V.

FT-ICR MS analysis of decorin glycosaminoglycan. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) experiments were performed in negative-ion mode with a 9.4 T Bruker Apex Ultra QeFTMS (Bruker Daltonics) fitted with an Apollo II dual source. Solutions of each decorin glycosaminoglycan fraction were introduced at a concentration of ~0.001 mg/mL in 50% aqueous MeOH. The sample solutions were infused at a rate of 7-14 μL per hour and were ionized by nanoelectrospray using a pulled fused silica-tip model FS360-75-15-d-20 (New Objective). Compositions from mass spectra where the monoisotopic peak was observed were assigned with mass accuracy of

6 ppm or better after external calibration. Isotope packets where monoisotopic peaks were not observed directly were extrapolated mathematically and assigned compositions with mass accuracy of 10 ppm or better. Tandem MS experiments were performed by mass selection of precursor ions and activation by CID in the hexapole collision cell of the Apex instrument with specified collision energies ranging from 8-20 V. 24-48 scans were signal averaged for each tandem MS experiment with average transient lengths of 0.75 s and 200,000 mass resolving power at m/z 400. Glycosidic bond cleavages were assigned with mass accuracy of less than 5 ppm with in-house developed software described below.

Automated Tandem MS Data Analysis. Overall search space of viable structures for each chain length is shown in Table 1. Analysis of over 1 million structures by manual examination is impractical. In-house software developed in the MATLAB (Mathworks) coding environment is used to automate the interpretation of tandem MS data. Modification and chain lengths are provided as user inputs after composition determination by accurate mass measurement. The software employs a genetic algorithm in combination with the sulfate modification limitations based on disaccharide analysis to drastically reduce overall search space of possible structures for unknown glycans. Tandem mass spectra of unknown structures are compared against theoretical structures using a closeness-of-fit function based on the number of matching glycosidic fragments, fragment intensities and overall depth of sequence coverage. Differentiation between 4-*O*- and 6-*O*-sulfo group modifications occurs when diagnostic cross-ring fragmentation is observed. Sulfo groups without diagnostic cross-ring fragments are not assigned a specific position and intentionally left ambiguous with a notation of S in annotated structures. Charge-state matching and isotope-pattern recognition were verified first using an automated charge-

spacing module and then by manual examination when interpreting FT-ICR-MS/MS data. A more comprehensive and detailed explanation of the algorithm and the software is provided in the doctoral thesis of J. Duan.[37]

4.5 CONCLUSIONS

Similar to the bikunin GAG chain, the porcine skin decorin GAG chain appears to also show a major structural motif with no variation within a short region comprising the six saccharide units at the chain's reducing end (Figure 4.6). The next 12 saccharides (residues 7-18) show subtle variability in both sulfation and uronic acid epimers. The remaining saccharides (residues 18-30) extending to the non-reducing end of the chain are enriched (~75%) in →4) IdoA (1→3) GalNAc4S (1→ repeating units. These flexible IdoA-rich domains are believed to be responsible for much of the protein-GAG interaction associated with decorins GAG-chain mediated activities. In summary, the heterogeneity in the decorin GAG chain structure is infrequent and occurs as a minor percentage of the overall sequence. These results should simplify future structure-activity relationship studies on decorin PGs.

It remains to be seen if the relatively invariant structures of the chondroitin sulfate PG, bikunin, and dermatan sulfate PG, decorin, extend to the heparan sulfate family. Heparan sulfates contain *N*-sulfo groups and both uronic acid epimers so that they are much more structurally complex. We estimate the theoretical structural variability of a single heparan sulfate chain occupying a specific site within a core protein to be > 10^{24}. The current study on decorin demonstrates that MS analysis is possible on GAG chains of dp 20-44 with 8-20 sulfo groups/GAG chain. Improved MS technology will be required for the longer and

more highly sulfated heparan sulfate GAG chains. Moreover, the routine determination of the C-5 uronic acid stereochemistry and the potential lability of heparan sulfate's *N*-sulfo groups will require extensive exploration. While progress continues to be made in GAG sequencing, additional improvements in separation technology and mass spectrometry will be required to sequence these incredibly complex family of biomacromolecules.

4.6 REFERENCES

(1) Hart, G. W.; Copeland, R. J. *Cell*. **2010**, 143, 672.

(2) Pilobello, K. T.; Mahal, L. K. *Curr. Opin. Chem. Biol*. **2007**, 11, 300.

(3) Tyers, M.; Mann, M. *Nature*. **2003**, 422, 6928.

(4) Neelamegham, S.; Mahal, L. K. *Curr. Opin. Chem. Biol*. **2016**, 40, 145.

(5) Vetr. H.; Gebhard, W. *Bio. Chem. Hoppe. Seyler*. **1990**, 371, 1185.

(6) Hascall, V. C.; Sajdera, S. W. *J. Biol. Chem*. **1970**, 245, 4920.

(7) Kitagawa, H.; Oyama, M.; Masayama, K.; Yamaguchi, Y.; Sugahara, K. *Glycobiology*. **1997**, 7, 1175.

(8) Fransson, L. A.; Belting, M., Jönsson, M.; Mani, K.; Moses, J.; Oldberg, A. *Matrix. Biol*. **2002**, 19, 367.

(9) Malmström, A.; Bartolini, B.; Thelin, M. A.; Pacheco, B.; Maccarana, M. *J. Histochem. Cytochem*. **2012**, 60, 916.

(10) Ly, M.; Leach, F. E. III.; Laremore, T. N.; Toida, T.; Amster, I. J.; Linhardt, R. J. *Nat. Chem. Biol*. **2011**, 7, 827.

(11) Kiani, C.; Chen, L.; Wu, Y. J.; Yee, A. J.; Yang, B. B. *Cell. Res*. **2002**, 12, 19.

(12) Chi, L.; Wolff, J. J.; Laremore, T. N.; Restaino, O. F.; Xie, J.; Schiraldi, C.; Toida, T.; Amster, I. J.; Linhardt, R. J. *J. Am. Chem. Soc*. **2008**, 130, 2617.

(13) Lamkin, E.; Cheng, G.; Calabro, A.; Hascall, V. C.; Joo, E. J.; Li, L.; Linhardt, R. J.; Lauer, M. E. *J. Biol. Chem*. **2015**, 290, 5156.

(14) Laremore, T. N,; Ly, M.; Zhang, Z.; Solakyildirim, K.; McCallum, S. A.; Owens, R. T.; Linhardt, R. J. *Biochem. J.* **2010**, 431, 199.

(15) Zhao, X.; Yang, B.; Solakylidirim, K.; Joo, E. J.; Toida, T.; Higashi, K.; Linhardt, R. J.; Li, L. *J. Biol. Chem*. **2013**, 288, 9226.

(16) Toyoda, H.; Kobayashi, S.; Sakamoto, S.; Toida, T.; Imanari, T. *Biol. Pharm. Bull.* **1993**, 16, 945.

(17) Yamada, S.; Oyama, M.; Kinugasa, H.; Nakagawa, T.; Kawasaki, T.; Nagasawa, S.; Khoo, K. H.; Morris, H. R.; Dell. A.; Sugahara, K. *Glycobiology*. **1995**, 5, 335.

(18) Zamfir, A.; Seidler, D. G.; Kresse, H.; Peter-Katalinic. *Glycobiology*. **2003**, 13, 733.

(19) Seo, N. S.; Hocking, A. M.; Höök, M.; McQuillan, D. J. *J. Biol. Chem*. **2005**, 280, 42774.

(20) Hocking, A. M.; Shinomura, T.; McQuillan, D. J. *Matrix. Biol*. **1998**, 17, 1.

(21) Seidler, D. G.; Dreier, R. *IUBMB Life*. **2008**, 60, 729.

(22) Järveläinen, H.; Sainio, A.; Wight, T. N. *Matrix. Biol*. **2015**, 43, 15.

(23) Gubbiotti, M. A.; Vallet, S. D.; Ricard-Blum, S.; Iozzo, R. V. *Matrix. Biol*. **2016**, 55, 7.

(24) Schaefer, L.; Tredup, C.; Gubbiotti, M. A.; Iozzo, R. V. *FEBS. J*. **2017**, 284, 10.

(25) Lin, Y. P.; Osburne, M. S.; Pereira, M. J.; Coburn, J.; Leong, M. *CRC Press*. **2016**, 86.

(26) Lin, Y.; L. Li, F. Zhang, R. J. Linhardt. Microbiol. Res. Submitted, **2017**

(27) Linhardt, R. J.; Hileman, R. E. *Gen. Pharmacol*. **1995**, 26, 443.

(28) Malavaki, C.; Mizumoto, S.; Karamanos, N.; Suqahara, K. *Connect. Tissue. Res*. **2008**, 49, 133.

(29) Pojasek, K.; Shriver, Z.; Kiley, P.; Venkataraman, G.; Sasisekharan, R. *Biochem. Bioph. Res. Co*. **2001**, 286, 343.

(30) Linhardt, R. J. *Curr. Protoc, Mol. Biol.* **2001**, Chapter 17: Unit 17. 13B.

(31) He, W.; Zhu, Y.; Shirke, A.; Sun, X.; Liu, J.; Gross, R. A.; Koffas, M. A. G.; Linhardt, R. J.; Li, M. *Appl. Microbiol. Biotechnol*. **2017**, 101, 6919.

(32) Zaia, J.; Li, X.Q.; Chan S.Y.; Costello C.E. *J. Am. Soc. Mass. Spectr.* **2003,** 14, 11.

(33) Bielik, A. M.; Zaia. *J Int. J. Mass Spectrom*. 2011, 305, 131.

(34) Leach, F. E.; Ly, M.; Laremore, T. N.; Wolff, J. J.; Perlow, J.; Linhardt, R. J.; Amster, I. J. *J. Am. Soc. Mass. Spectr.* **2012**, 23, 1488.

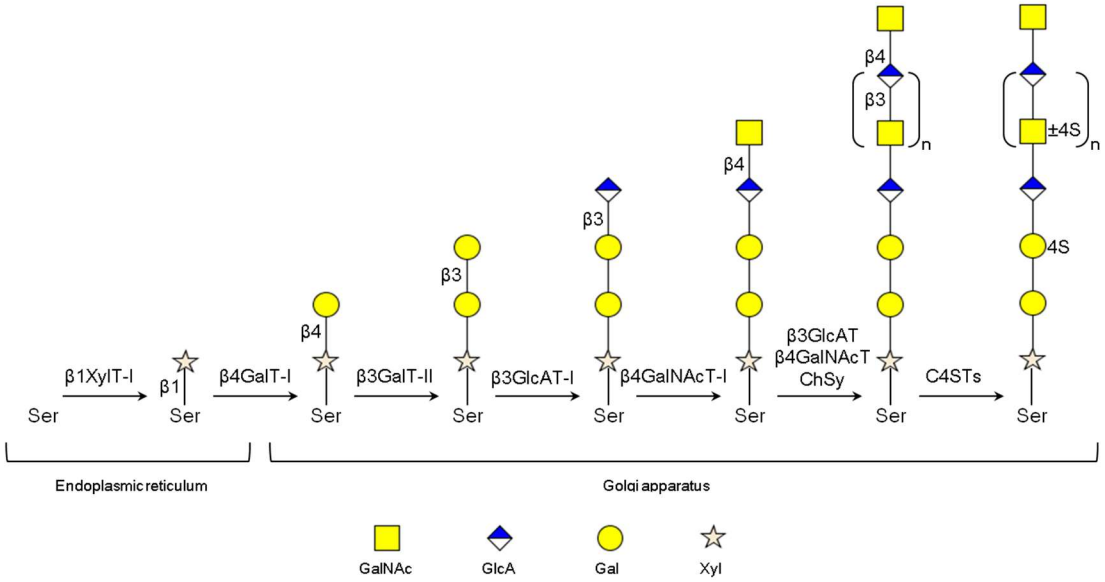
(35) Kailemia, M. J.; Patel, A. B.; Johnson, D. T.; Li, L.; Linhardt, R. J.; Amster, I. J. *Eur. J. Mass. Spectrom*. **2015**, 21, 275.

(36) Agyekum, I; Zong, C. L; Boons, G. J; Amster, I. J. *J. Am. Soc. Mass. Spectr.* **2017,** 28, 9.

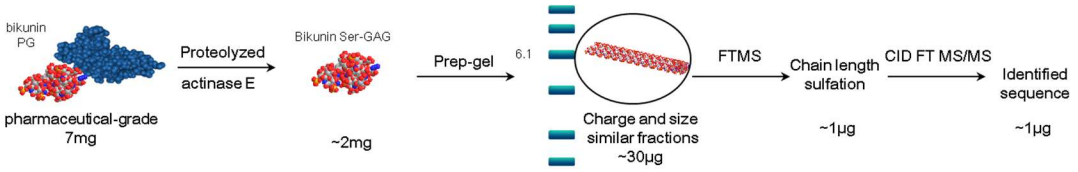
(37) Williams, A.; He, W.; Cress, B. F.; Liu, X.; Alexandria, J.; Yoshizawa, H.; Nishimuram, K.; Toida, T.; Koffas, M.; Linhardt, R. J. *Biotechnol. J.* **2017**, 12, 1700239.

(38) Duan, J. Doctoral Thesis, Department of Chemistry, University of Georgia, **2017.**

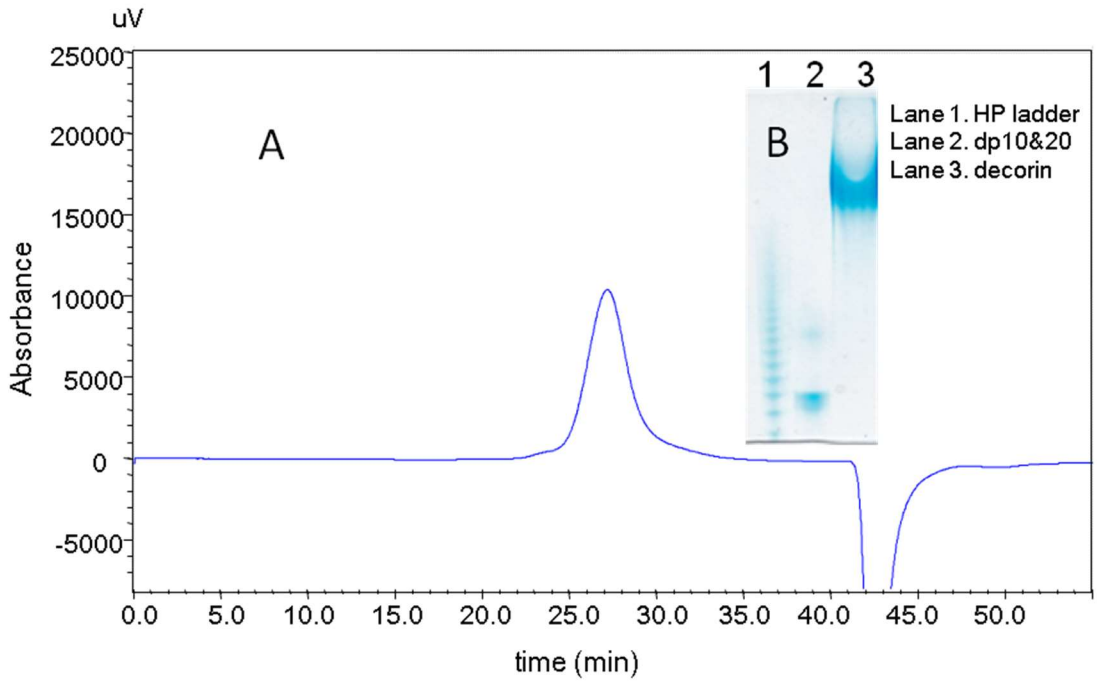
4.7 SUPPLEMENTAL FIGURES AND TABLES



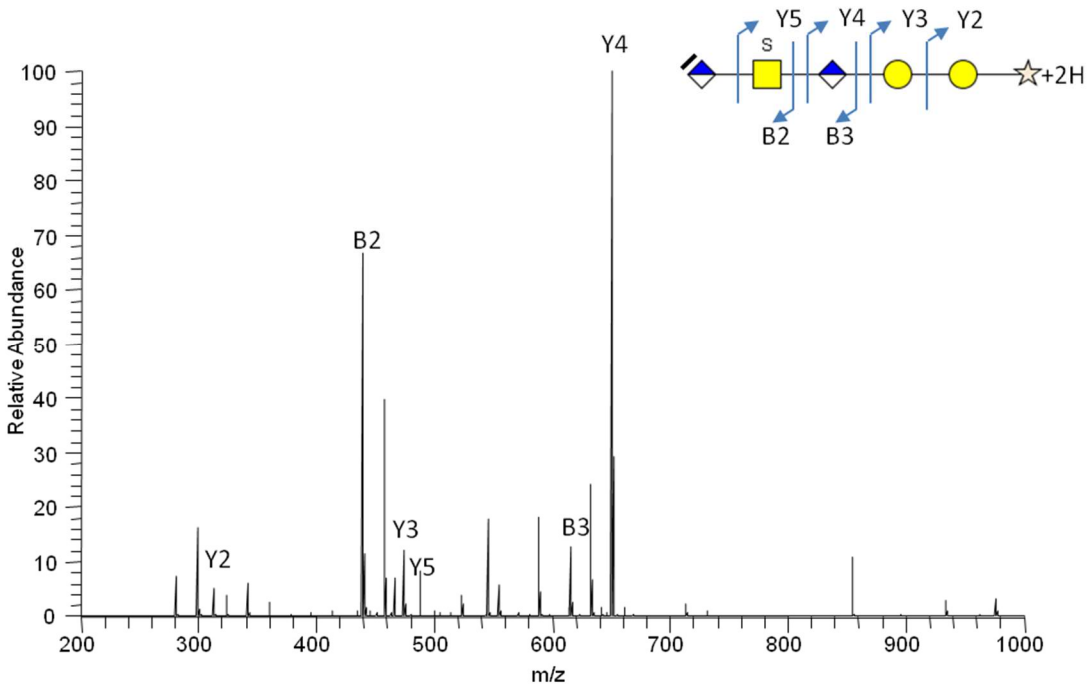
Supplementary Figure 4.1. Biosynthetic pathway for chondroitin sulfate A GAG. The GAG, on a serine residue of the core protein, is synthesized in a pathway that begins in the endoplasmic reticulum and concludes in the Golgi apparatus. The biosynthetic enzymes are: β1XylT-I, β-xylosyl transferase I; β4GalT-I, β-4-galactosyl transferase I; β3GalT-II, β-3-galactosyl transferase II; β3GlcAT-I, β-3-glucuronosyl transferase I; β4GalNAcT-I, β-4-N-acetyl galactosaminyl transferase I; β3GlcAT, β-3-glucuronosyl transferase; β4GalNAcT, β-4-N-acetyl galactosaminyl transferase; ChSy, chondroitin synthases; C4STs, galactosyl 4-O-sulfo transferase and N-acetyl galactosaminyl-4-O-sulfotransferase.



Supplementary Figure 4.2. Bikunin flow chart for solving structure. Pharmaceutical-grade bikunin proteoglycan was proteolyzed by actinase E to obtain glycosaminoglycan. GAG was then fractionated by continuous elution preparative PAGE to get charge and size similar chains. Chains were identified by FTMS and CID FT MS/MS.



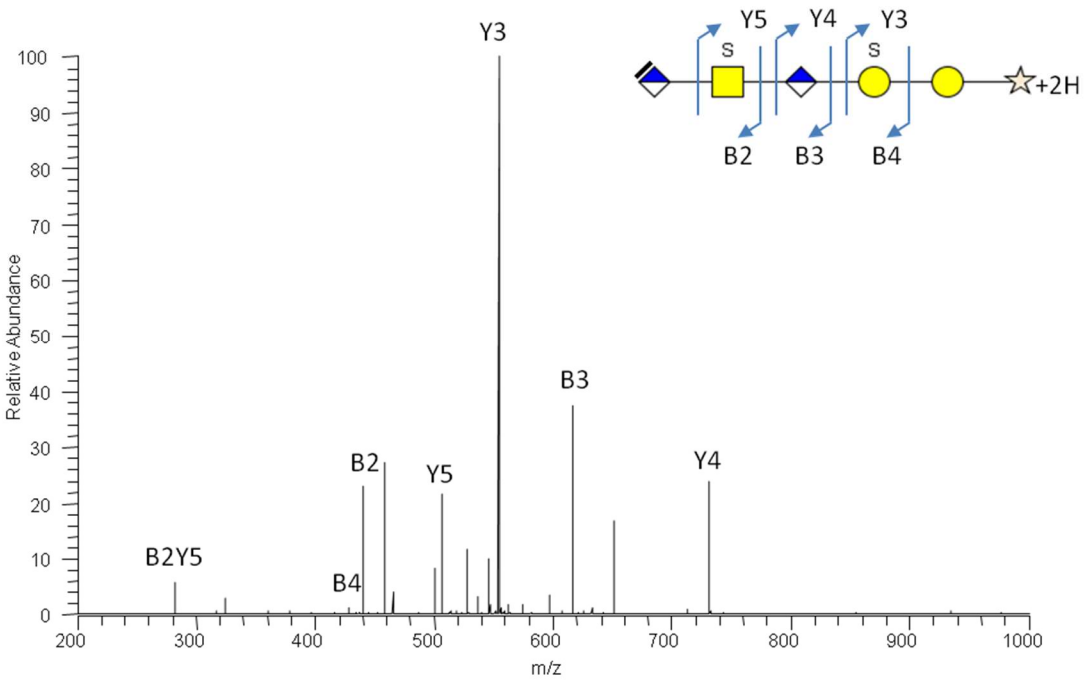
Supplementary Figure 4.3. Molecular weight analysis of decorin glycosaminoglycan. A. GPC-HPLC analysis of porcine decorin GAG. The column was calibrated using heparin standard of defined molecular weight. B. 15% PAGE analysis of decorin GAG using heparin ladder and dp10&20 as standards. The molecular weight was calculated by UN SCAN IT software.



cleavage	m/z	Rel. Inten.	Ion	Error (ppm)	M_{exp}	M_{theor}
B2	440.0498	67.17	ΔHexA1HexNAc1S1-B	-0.2	441.0576	441.0577
B3	616.0811	12.85	ΔHexA1HexNAc1HexA1S1-B	-1.5	617.0889	617.0898
Y2	313.1131	7.68	Y-Hex1Pen1	-1.3	314.1209	314.1213
Y3	475.1656	18.12	Y-Hex2Pen1	-1.5	476.1734	476.1741
Y4	651.1971	100	Y-HexA1Hex2Pen1	-2.0	652.2049	652.2062
Y5	934.2318	5.27	Y-HexNAc1HexA1Hex2Pen1S1	-3.0	935.2396	935.2424
Y5	466.6123	7.24	Y-HexNAc1HexA1Hex2Pen1S1	-2.4	935.2402	935.2424
Y5-S	854.2758	11.01	Y-HexNAc1HexA1Hex2Pen1	-2.3	855.2836	855.2856
M	545.6234	7.16	ΔHexA1HexNAc1HexA1Hex2Pen1S1	-1.4	1093.2624	1093.2639

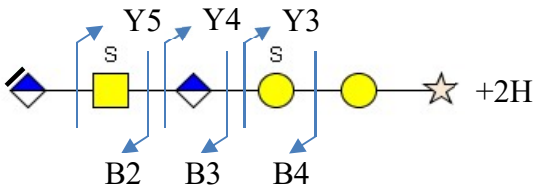
Supplementary Figure 4.4. FTMS analysis of linkage region. Negative-ion mode FTMS showed the hexasaccharide with single composition ΔUA-GalNAc4S-GlcA-Gal-Gal-Xylitol was most abundant. CID-FT MS/MS analysis of the hexasaccharide with m/z 545.6242. Assignment of fragment ions identified include B2, B3, Y2, Y3, Y4, Y5 which

afforded the uniform composition of the linkage region with a xylitol replace of xylose on the reducing end. Mass spectra were acquired on the Orbitrap FT MS.

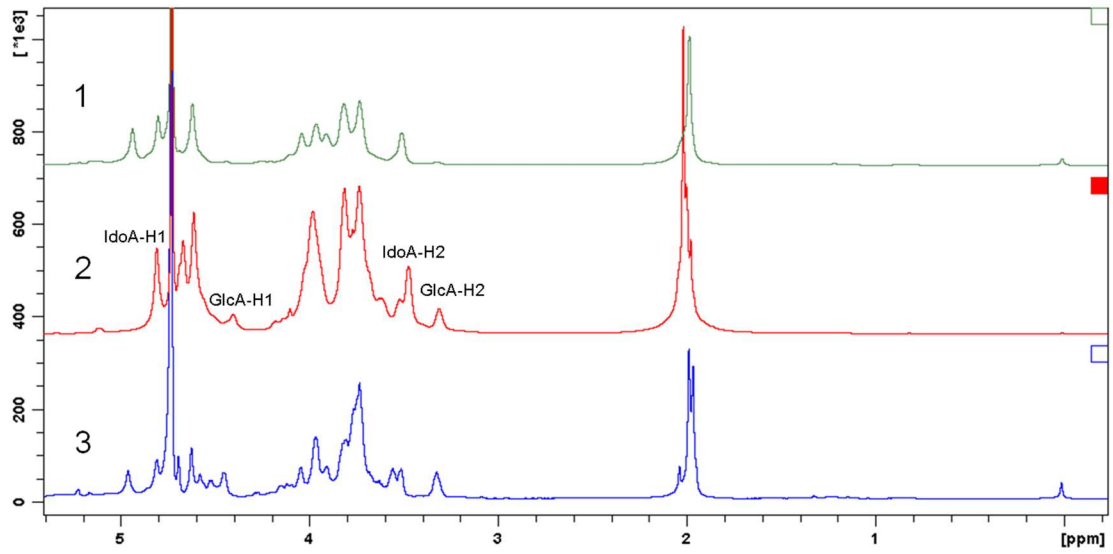


cleavage	m/z	Rel. Inten.	Ion	Error (ppm)	M_{exp}	M_{theor}
B2	440.0498	23.71	ΔHexA1HexNAc1S1-B	-0.2	441.0576	441.0577
B3	616.0814	38.32	ΔHexA1HexNAc1HexA1S1-B	-1.0	617.0892	617.0898
B4	428.5414	1.14	ΔHexA1HexNAc1HexA1Hex1S2-B	-1.2	859.0984	859.0994
Y3	555.1226	100	Y-Hex2Pen1S1	-0.9	556.1304	556.1309
Y4	731.1541	24.06	Y-HexA1Hex2Pen1S1	-1.5	732.1619	732.163
Y4-S	651.1975	16.79	Y-HexA1Hex2Pen1	-1.4	652.2053	652.2062

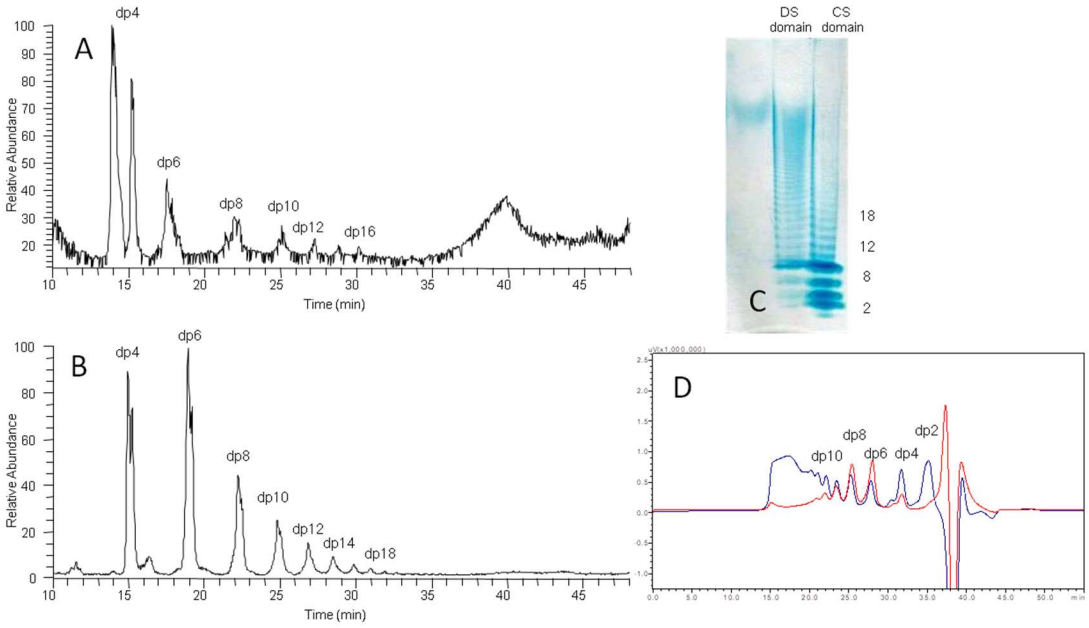
Y5	506.5914	21.74	Y-HexNAc1HexA1Hex2Pen1S2	-0.8	1015.1984	1015.1992
B2Y5	282.0283	5.7	ΔHexNAc1S1	-0.4	283.0361	283.0362



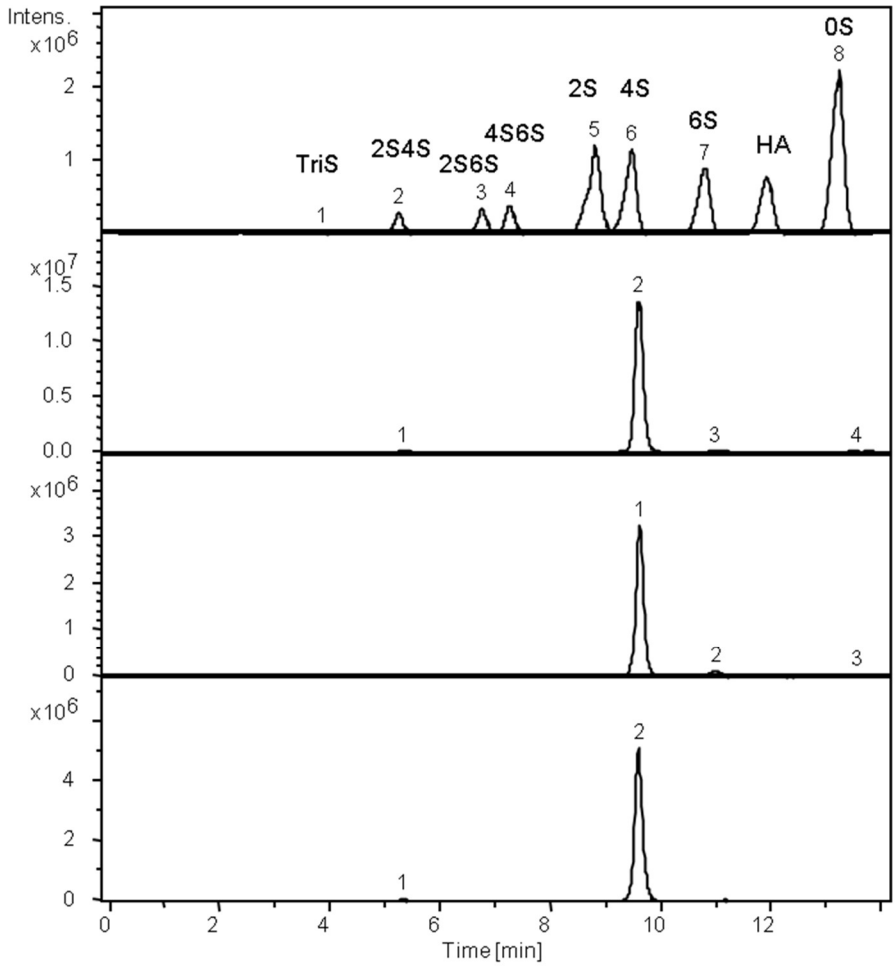
Supplementary Figure 4.5. FTMS analysis of linkage region. Negative-ion mode FTMS showed the hexasaccharide with 2 sulfation ΔUA-GalNAc4S-GlcA-Gal4S-Gal-Xylitol also existed. CID-FT MS/MS analysis of the hexasaccharide with m/z 585.6025. Assignment of fragment ions identified include B2, B3, B4, B2Y5, Y2, Y3, Y4 which afforded the uniform composition of the linkage region with a xylitol replace of xylose on the reducing end. Mass spectra were acquired on the Orbitrap FT MS.



Supplementary Figure 4.6. ^{1}H-NMR analysis of porcine skin decorin GAG performed in D_2O at 600 MHz. 1. B-type domain; 2. decorin GAG; 3. AC-type domain. The ratio of the IdoA-H1 and GlcA-H1 were calculated as 3:1

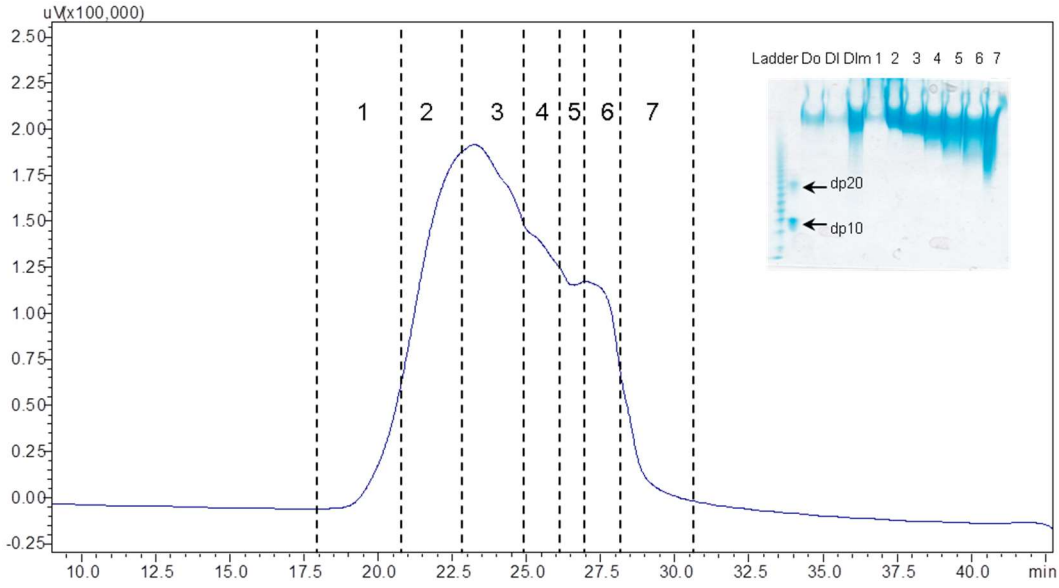


Supplementary Figure 4.7. Decorin domain analysis. A&B, TIC of HILIC MS analysis of porcine skin decorin. A, chondroitinase AC treatment converted AC-type domain products primary disaccharides and intact B-type domains. B, chondroitinase B treatment converted B-type disaccharides and intact AC-type domains of dp6-dp20 and linkage region. C, 15% PAGE analysis of AC-type and B-type domain. D, Size exclusion chromatography on HPLC with refractive index detector, red: AC-type domain; blue: B-type domain.

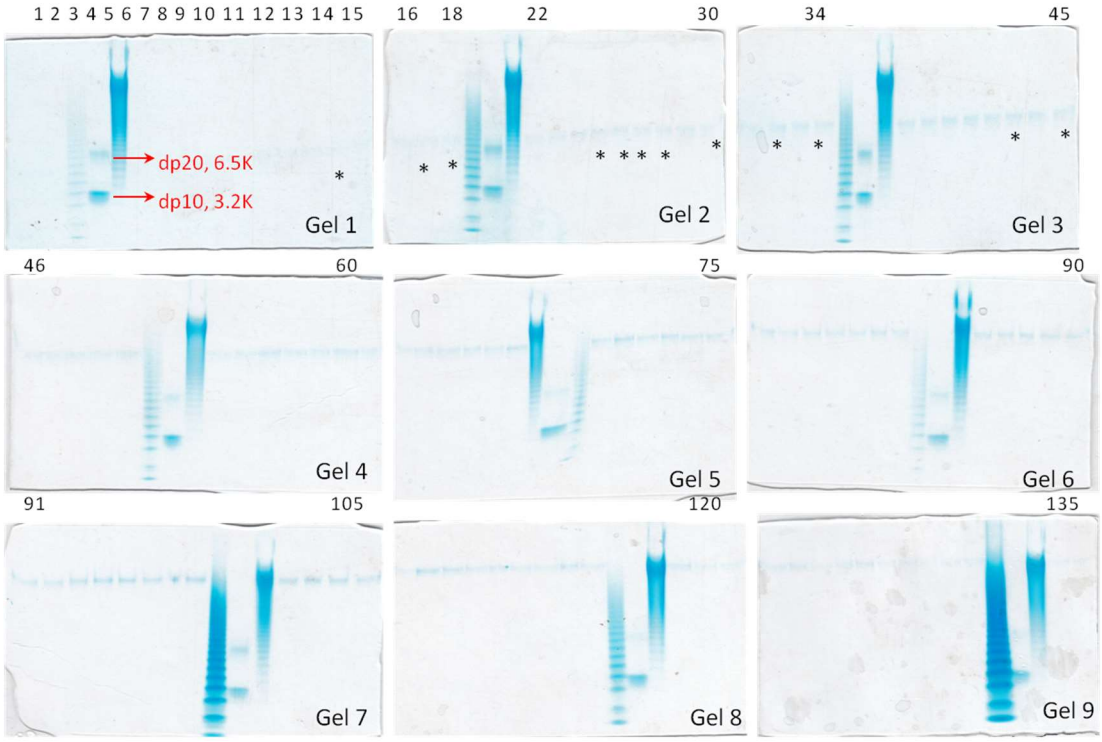


	Porcine decorin GAG		
Disaccharides	Original	AC-type domain	B-type domain
2S4S	2.0	0	4.7
4S	96.8	93.2	95.3
6S	1.0	6.0	0
0S	0.2	0.8	0

Supplementary Figure 4.8. Disaccharides compositional analysis



Supplementary Figure 4.9. Decorin of low molecular weight and medium charge fractions were separated on HPLC with refractive index detector. Samples were cut into 7 fractions based on time. Each fraction was visualized by 15% PAGE with Alcian blue stain labeled from 1 to 7. Heparin ladder (lane 1 from left) and dp10&20 (lane 2 from left) were used as molecular weight standard.

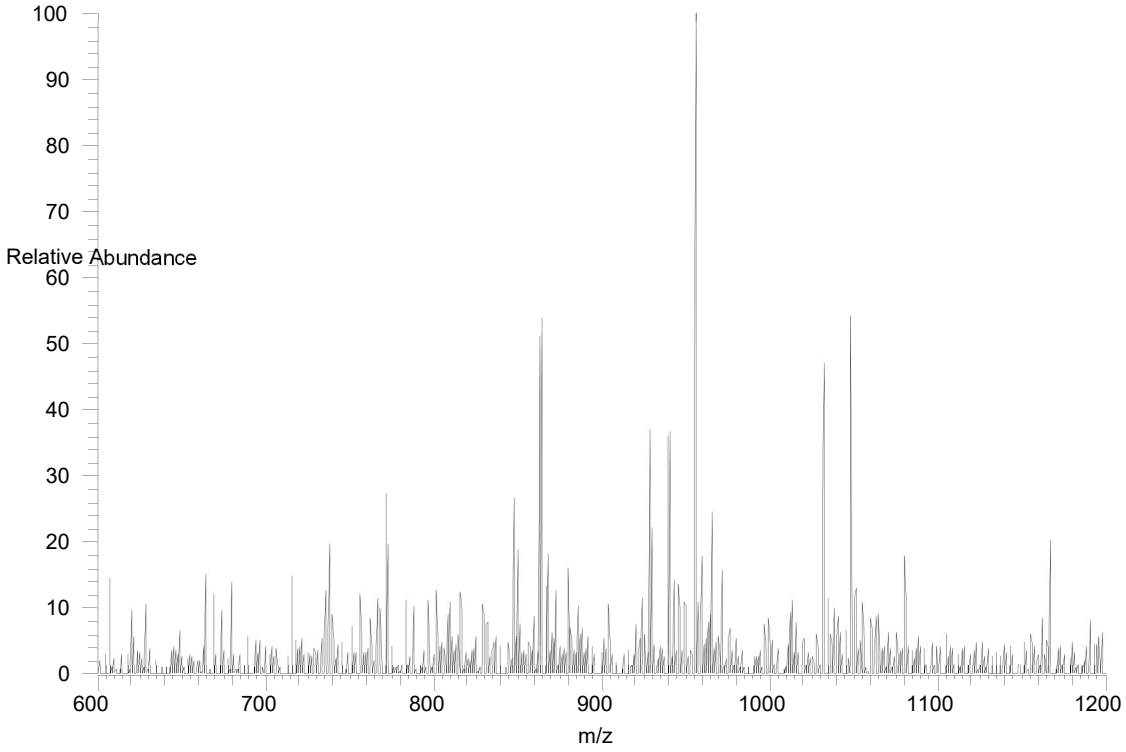


Supplementary Figure 4.10. Analytical PAGE analysis of decorin GAG. 15% PAGE numbered as Gel 1-9 allowed Molecular weight calculation of gel bands. Heparin ladder and dp10&20 were used as markers in each gel. Starred bands (*) analyzed by FTMS or FT-ICR MS.

Gel	Lane	Fraction	M_R by PAGE	Gel	Lane	Fraction	M_R by PAGE	Gel	Lane	Fraction	M_R by PAGE
1	1	25	3.0	4	46	61	9.0	7	91	97	20.8
	2	26	3.0		47	62	9.4		92	98	21.2
	6	27	3.2		48	63	9.8		93	99	21.1
	7	28	3.2		49	64	10.3		94	100	21.3
	8	29	3.3		50	65	10.4		95	101	21.6
	9	30	3.4		54	66	10.9		96	102	21.8
	10	31	3.5		55	67	11.3		97	103	21.9
	11	32	3.7		56	68	11.5		98	104	21.9
	12	33	3.7		57	69	11.9		102	105	22.1
	13	34	3.9		58	70	12.1		103	106	22.2
	14	**35**	**4.2**		59	71	12.5		104	107	22.4
	15	36	4.4		60	72	12.8		105	108	22.3
2	16	37	4.6	5	61	73	13.0	8	106	109	23.0
	17	**38**	**4.7**		62	74	12.9		107	110	23.3
	18	**39**	**5.0**		63	75	13.1		108	111	23.4
	22	40	5.0		64	76	13.2		109	112	23.8
	23	41	5.2		65	77	13.8		110	113	24.0
	24	42	**5.2**		66	78	14.2		111	114	24.1
	25	**43**	**5.5**		70	79	15.1		112	115	24.2
	26	**44**	**5.8**		71	80	15.6		113	116	24.5
	27	**45**	**6.1**		72	81	16.2		114	117	24.9
	28	**46**	**6.3**		73	82	16.7		118	118	25.1

	29	47	6.5		74	83	17.2		119	119	25.3
	30	**48**	**6.6**		75	84	17.4		120	120	25.6
3	31	49	6.8	6	76	85	17.6	9	121	121	26.6
	32	**50**	**7.0**		77	86	17.8		122	122	27.3
	33	51	7.0		78	87	17.9		123	123	28.2
	34	**52**	**7.2**		79	88	18.1		124	124	29.2
	38	53	7.5		80	89	18.4		125	125	30.0
	39	54	7.7		81	90	18.5		126	126	30.3
	40	55	7.7		82	91	18.8		127	127	30.5
	41	56	7.8		86	92	19.0		128	128	30.3
	42	57	8.0		87	93	19.2		129	129	30.6
	43	**58**	**7.9**		88	94	19.2		130	130	29.8
	44	59	8.3		89	95	19.5		134	134	30.7
	45	**60**	**8.7**		90	96	19.7		135	135	31.9

Supplementary Table 4.1. M_R-values of fractions #25-135, based on molecular mass densitometry analysis against heparin ladder standards. Their respective lane loadings from Supplementary Figure 4.4 are listed with rows in bold analyzed by FTMS or FT-ICR MS.

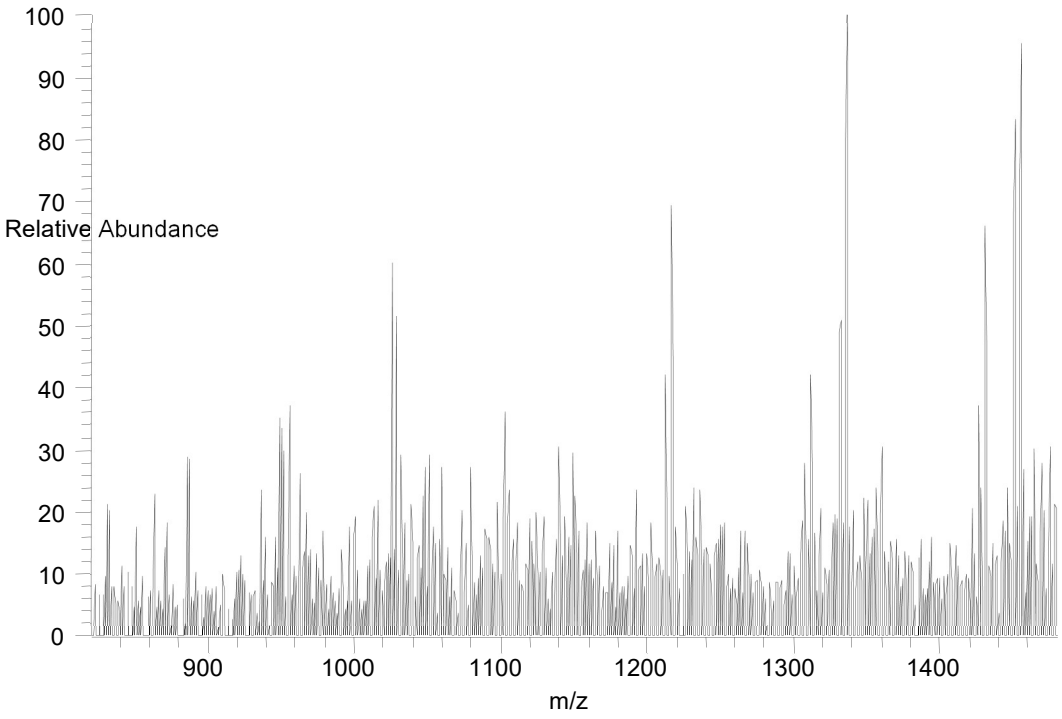


Supplementary Figure 4.11. FTMS of fraction 38 (4.7 kDa by PAGE) corresponding to 18 composition identifications listed in Supplementary Table 4.2.

m/z	z	M_{Exp}	M_{Theor}	Accuracy (ppm)	Chain
664.3151	5	3326.6145	3326.6592	-13.4	dp16-5S
830.6471	4	3326.6196	3326.6592	-11.9	dp16-5S
680.3069	5	3406.5735	3406.616	-12.5	dp16-6S
850.6389	4	3406.5868	3406.616	-8.6	dp16-6S
925.4268	4	3705.7384	3705.7707	-8.7	dp18-5S
756.1383	5	3785.7305	3785.7275	0.8	dp18-6S
772.1198	5	3865.638	3865.6843	-12.0	dp18-7S
965.4002	4	3865.632	3865.6843	-13.5	dp18-7S
788.1138	5	3945.608	3945.6411	-8.4	dp18-8S
847.941	5	4244.744	4244.7958	-12.2	dp20-7S
1060.1782	4	4244.744	4244.7958	-12.2	dp20-7S
863.9322	5	4324.7	4324.7526	-12.2	dp20-8S
1080.1694	4	4324.7088	4324.7526	-10.1	dp20-8S
879.9294	5	4404.686	4404.7094	-5.3	dp20-9S
920.5432	5	4607.755	4607.7888	-7.3	dp21-9S
923.7668	5	4623.873	4623.9073	-7.4	dp22-7S
939.7535	5	4703.8065	4703.8641	-12.2	dp22-8S
796.2911	6	4783.7934	4783.8209	-5.7	dp22-9S
955.7447	5	4783.7625	4783.8209	-12.2	dp22-9S
971.7425	5	4863.7515	4863.7777	-5.4	dp22-10S
1012.3568	5	5066.823	5066.8571	-6.7	dp23-10S
1015.5781	5	5082.9295	5082.9756	-9.1	dp24-8S

859.4618	6	5162.8176	5162.9324	-22.2	dp24-9S
1031.5688	5	5162.883	5162.9324	-9.6	dp24-9S
1047.5594	5	5242.836	5242.8892	-10.1	dp24-10S
872.799	6	5242.8408	5242.8892	-9.2	dp24-10S

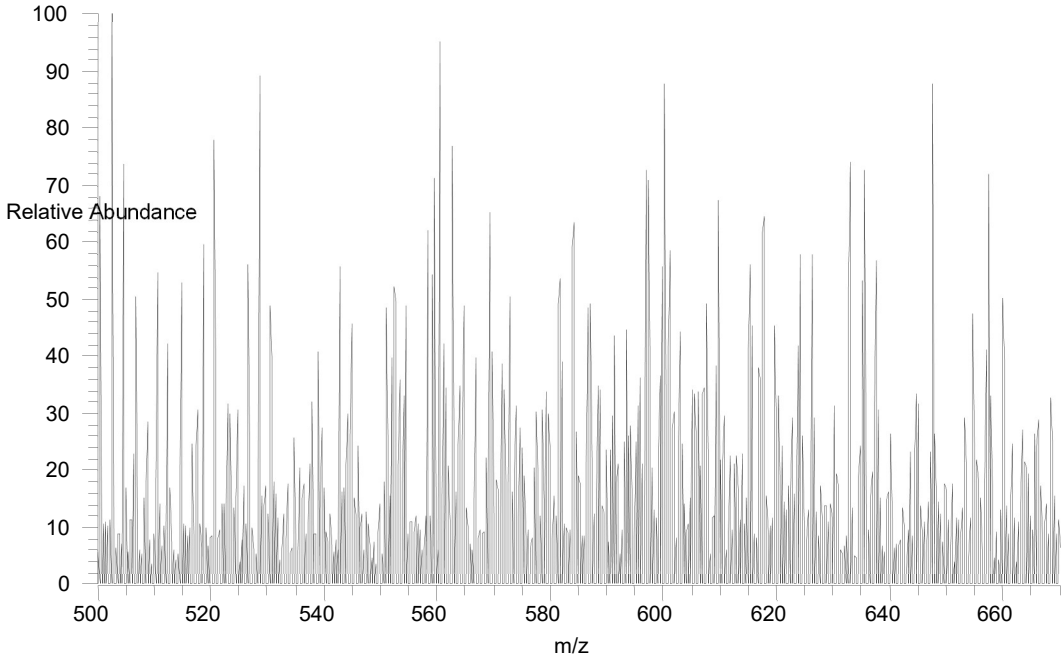
Supplementary Table 4.2. Accurate mass measurements for 18 compositions identified by FTMS of f38 (4.7 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.12. FTMS of fraction 44 (5.8 kDa by PAGE) corresponding to 13 composition identifications listed in Supplementary Table 4.3.

m/z	z	Mexp	Mtheor	Accuracy(ppm)	Chain
863.9402	5	4324.74	4324.7526	-2.9	dp20-8S
955.7525	5	4783.8015	4783.8209	-4.1	dp22-9S
1047.5665	5	5242.8715	5242.8892	-3.4	dp24-10S
1309.7095	4	5242.8692	5242.8892	-3.8	dp24-10S
1335.1958	4	5344.8144	5344.828	-2.5	dp24-11S(1Na)
1360.4964	4	5446.0168	5445.9686	8.9	dp25-10S
935.9942	6	5622.012	5622.0007	2.0	dp26-10S
949.3206	6	5701.9704	5701.9575	2.3	dp26-11S
1430.032	4	5724.1592	5723.9395	38.4	dp26-11S(1Na)
1450.0028	4	5804.0424	5803.8963	25.2	dp26-12S(1Na)
1012.5072	6	6081.09	6081.069	3.5	dp28-11S
1025.8199	6	6160.9662	6161.0258	-9.7	dp28-12S
1039.1601	6	6241.0074	6240.9826	4.0	dp28-13S
1347.6483	5	6743.2805	6743.2167	9.5	dp31-12S
1102.3387	6	6620.079	6620.0941	-2.3	dp30-13S

Supplementary Table 4.3. Accurate mass measurements for 13 compositions identified by FTMS of f44 (5.8 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.13. FTMS of fraction 45 (6.1 kDa by PAGE) corresponding to 14 composition identifications listed in Supplementary Table 4.4.

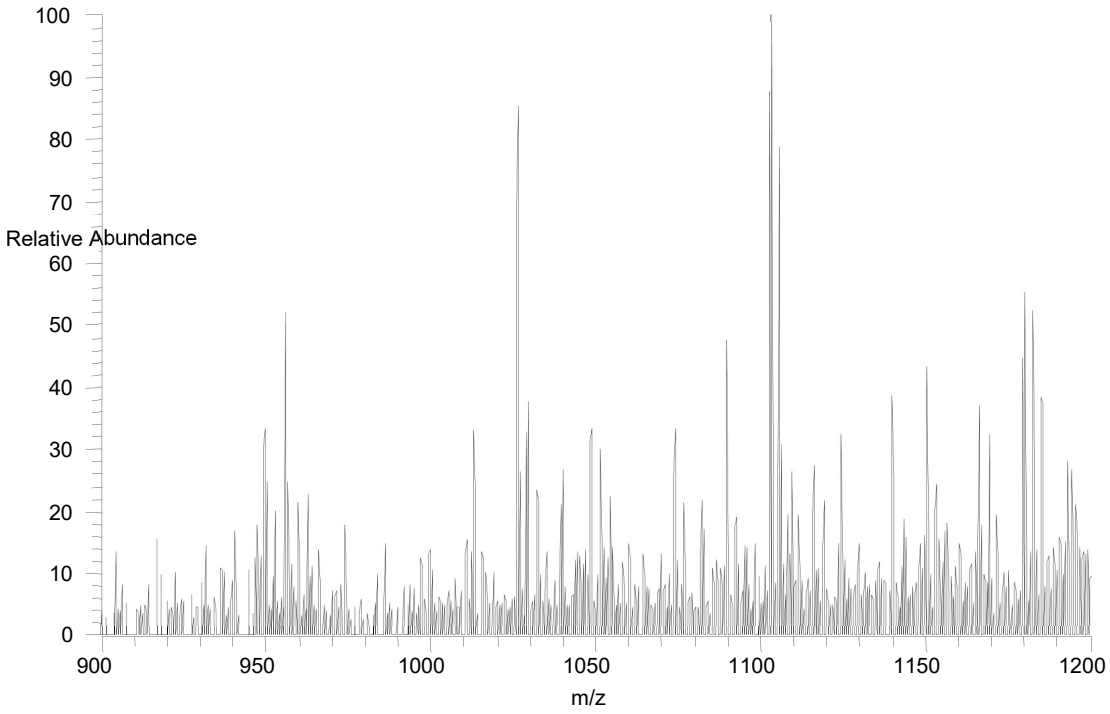
m/z	z	M_{Exp}	M_{Ther}	Accuracy (ppm)	Chain
629.9385	6	3785.6778	3785.7275	-13.1	dp18-6S
616.8071	7	4324.7043	4324.7526	-11.2	dp20-8S
530.5238	9	4783.7844	4783.8209	-7.6	dp22-9S
596.9663	8	4783.7928	4783.8209	-5.9	dp22-9S
599.7127	8	4805.764	4805.8029	-8.1	dp22-9S (1Na)
572.6507	9	5162.9265	5162.9324	-1.1	dp24-9S
644.3566	8	5162.9152	5162.9324	-3.3	dp24-9S
647.1002	8	5184.864	5184.9144	-9.7	dp24-9S (1Na)
523.277	10	5242.848	5242.8892	-7.9	dp24-10S
581.531	9	5242.8492	5242.8892	-7.6	dp24-10S
583.9734	9	5264.8308	5264.8712	-7.7	dp24-10S(1Na)
657.0959	8	5264.8296	5264.8712	-7.9	dp24-10S(1Na)
623.6557	9	5621.9715	5622.0007	-5.2	dp26-10S
626.0969	9	5643.9423	5643.9827	-7.2	dp26-10S(1Na)
704.4905	8	5643.9864	5643.9827	0.7	dp26-10S(1Na)
517.3488	11	5701.9226	5701.9575	-6.1	dp26-11S
632.5397	9	5701.9275	5701.9575	-5.3	dp26-11S
571.3801	10	5723.879	5723.9395	-10.6	dp26-11S(1Na)
634.9851	9	5723.9361	5723.9395	-0.6	dp26-11S(1Na)
637.4203	9	5745.8529	5745.9215	-11.9	dp26-11S(2Na)
524.6188	11	5781.8926	5781.9143	-3.8	dp26-12S
591.1002	10	5921.08	5921.1554	-12.7	dp28-9S
599.0985	10	6001.063	6001.1122	-8.2	dp28-10S

665.7739	9	6001.0353	6001.1122	-12.8	dp28-10S
668.22	9	6023.0502	6023.0942	-7.3	dp28-10S(1Na)
551.8125	11	6081.0233	6081.069	-7.5	dp28-11S
607.0913	10	6080.991	6081.069	-12.8	dp28-11S
674.6632	9	6081.039	6081.069	-4.9	dp28-11S
609.2931	10	6103.009	6103.051	-6.9	dp28-11S(1Na)
677.1048	9	6103.0134	6103.051	-6.2	dp28-11S(1Na)

Supplementary Table 4.4. Accurate mass measurements for 14 compositions identified by FTMS of f45 (6.1 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.

m/z	z	M_{Exp}	M_{Ther}	Accuracy (ppm)	Chain
559.0781	11	6160.9449	6161.0258	-13.1	dp28-12S
615.1015	10	6161.093	6161.0258	10.9	dp28-12S
683.5448	9	6160.9734	6161.0258	-8.5	dp28-12S
617.2876	10	6182.954	6183.0078	-8.7	dp28-12S(1Na)
685.9911	9	6182.9901	6183.0078	-2.9	dp28-12S(1Na)
543.9984	12	6540.0744	6540.1373	-9.6	dp30-12S
593.5464	11	6540.0962	6540.1373	-6.3	dp30-12S
653.0011	10	6540.089	6540.1373	-7.4	dp30-12S
595.5433	11	6562.0621	6562.1193	-8.7	dp30-12S(1Na)
600.8154	11	6620.0552	6620.0941	-5.9	dp30-13S
663.1957	10	6642.035	6642.0761	-6.2	dp30-13S(1Na)

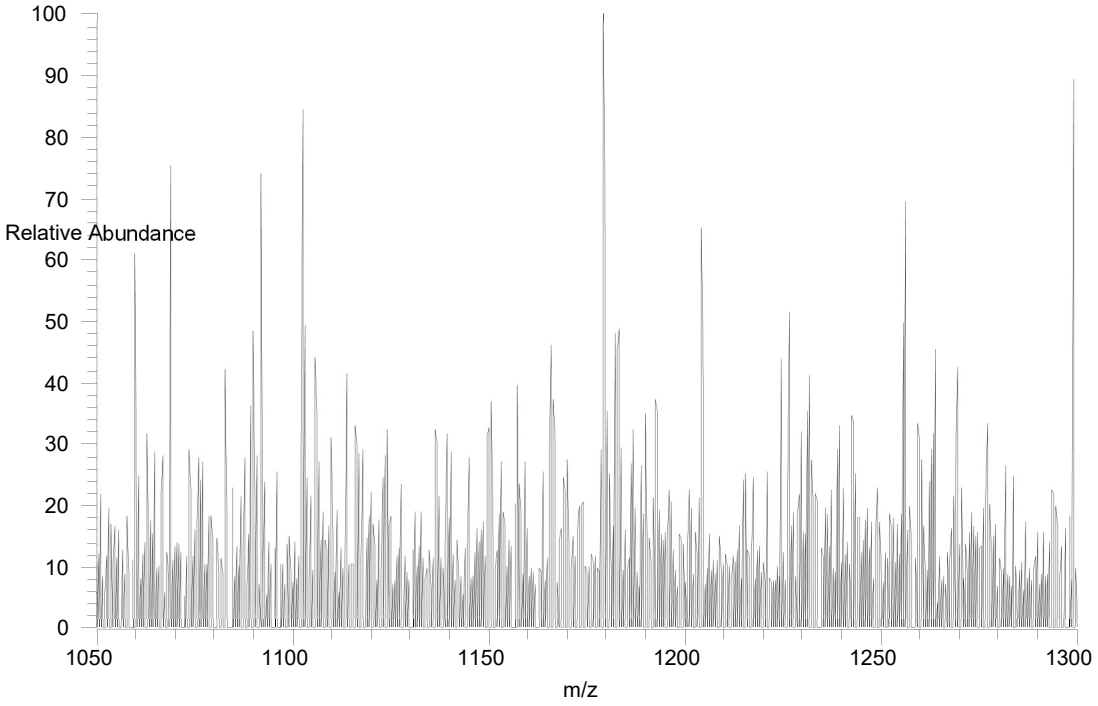
Supplementary Table 4.4. (continued)



Supplementary Figure 4.14. FTMS of fraction 48 (6.6 kDa by PAGE) corresponding to 18 composition identifications listed in Supplementary Table 4.5.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
955.7608	5	4783.843	4783.8209	4.6	dp22-9S
1031.5631	5	5162.8545	5162.9324	-15.1	dp24-9S
1047.5714	5	5242.896	5242.8892	1.3	dp24-10S
1063.5598	5	5322.838	5322.846	-1.5	dp24-11S
1123.4021	5	5622.0495	5622.0007	8.7	dp26-10S
1139.3805	5	5701.9415	5701.9575	-2.8	dp26-11S
949.3163	6	5701.9446	5701.9575	-2.3	dp26-11S
996.4883	6	5984.9766	5984.9937	-2.9	dp27-12S
1012.5004	6	6081.0492	6081.069	-3.3	dp28-11S
1025.8328	6	6161.0436	6161.0258	2.9	dp28-12S
1039.1504	6	6240.9492	6240.9826	-5.4	dp28-13S
1072.9938	6	6444.0096	6444.062	-8.1	dp29-13S
1089.0069	6	6540.0882	6540.1373	-7.5	dp30-12S
1102.3339	6	6620.0502	6620.0941	-6.6	dp30-13S
1115.6672	6	6700.05	6700.0509	-0.1	dp30-14S
1149.5206	6	6903.1704	6903.1303	5.8	dp31-14S
1165.5281	6	6999.2154	6999.2056	1.4	dp32-13S
1178.8638	6	7079.2296	7079.1624	9.5	dp32-14S
1192.1844	6	7159.1532	7159.1192	4.7	dp32-15S

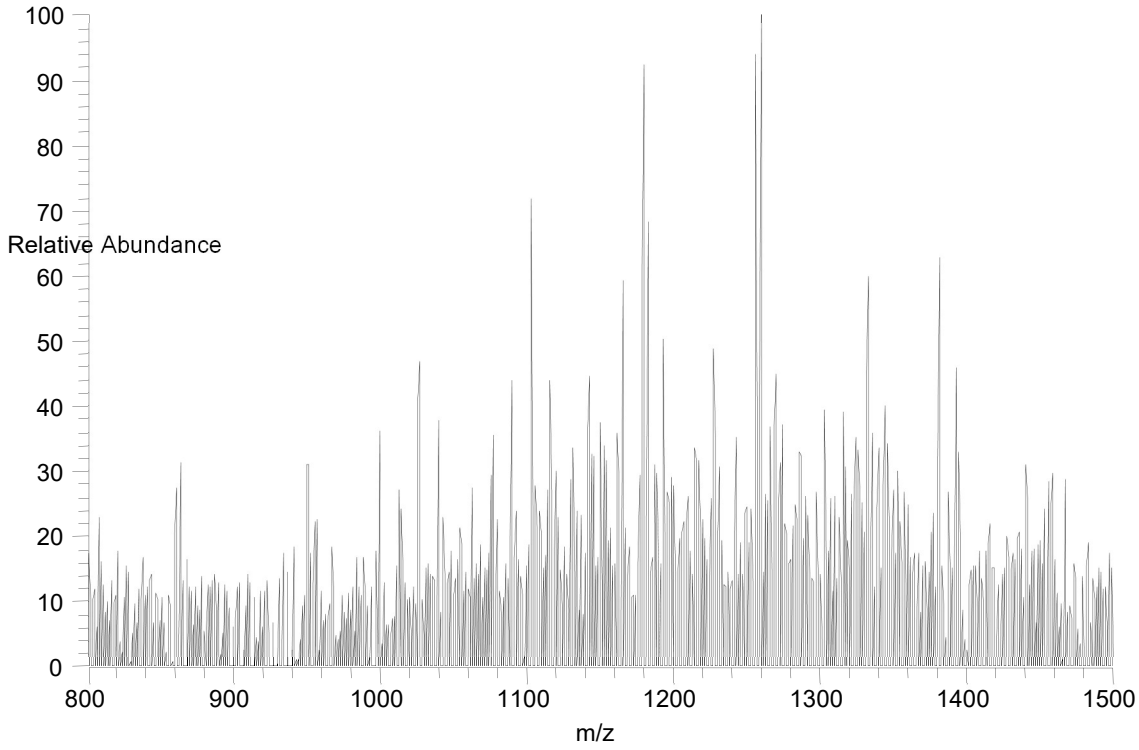
Supplementary Table 4.5. Accurate mass measurements for 18 compositions identified by FTMS of f48 (6.6 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.15. FTMS of fraction 50 (7.0 kDa by PAGE) corresponding to 10 composition identifications listed in Supplementary Table 4.6.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
1025.8294	6	6161.0232	6161.0258	-0.4	dp28-12S
1089.0115	6	6540.1158	6540.1373	-3.3	dp30-12S
1102.3429	6	6620.1042	6620.0941	1.5	dp30-13S
1165.5259	6	6999.2022	6999.2056	-0.5	dp32-13S
1178.8570	6	7079.1888	7079.1624	3.7	dp32-14S
1182.5216	6	7101.1764	7101.1444	4.5	dp32-14S (Na)
1226.0280	6	7362.2148	7362.1986	2.2	dp33-15S
1242.0315	6	7458.2358	7458.2739	-5.1	dp34-14S
1255.3619	6	7538.2182	7538.2307	-1.7	dp34-15S
1268.6915	6	7618.1958	7618.1875	1.1	dp34-16S

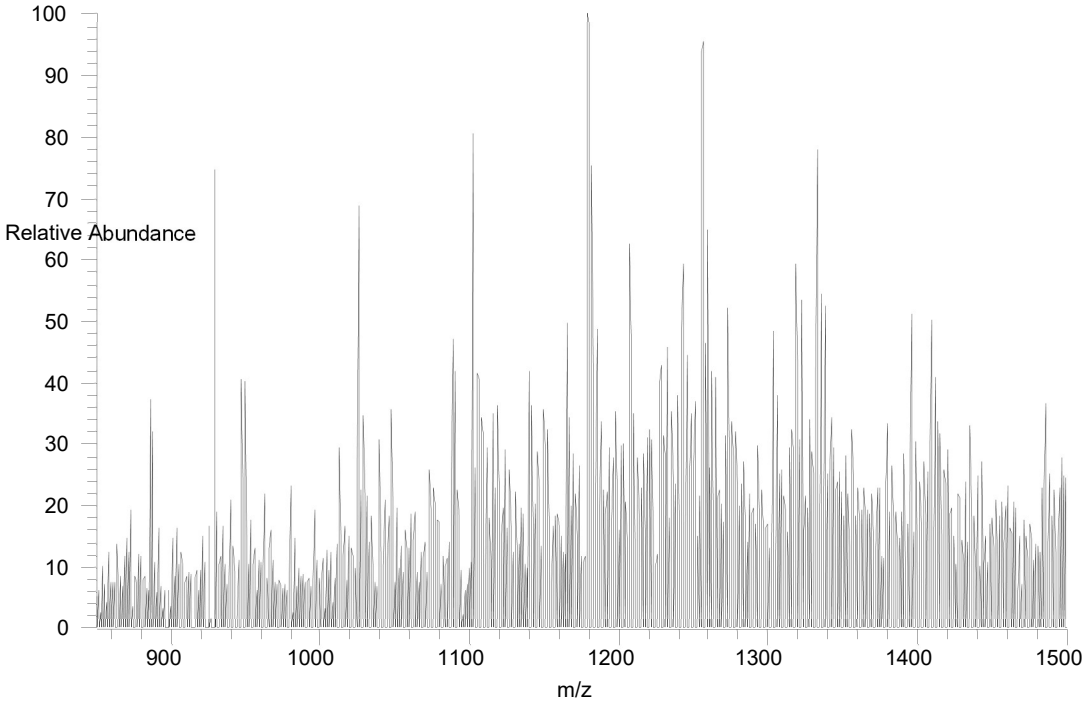
Supplementary Table 4.6. Accurate mass measurements for 10 compositions identified by FTMS of f50 (7.0 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.16. FTMS of fraction 52 (7.2 kDa by PAGE) corresponding to 17 composition identifications listed in Supplementary Table 4.7.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm	Chain
955.7456	5	4783.767	4783.8209	-11.3	dp22-9S
1047.5538	5	5242.808	5242.8892	-15.5	dp24-10S
1012.4895	6	6080.9838	6081.069	-14.0	dp28-11S
1025.8153	6	6160.9386	6161.0258	-14.2	dp28-12S
1039.1415	6	6240.8958	6240.9826	-13.9	dp28-13S
1089.0011	6	6540.0534	6540.1373	-12.8	dp30-12S
1102.3261	6	6620.0034	6620.0941	-13.7	dp30-13S
1149.4968	6	6903.0276	6903.1303	-14.9	dp31-14S
1165.5067	6	6999.087	6999.2056	-16.9	dp32-13S
1178.8433	6	7079.1066	7079.1624	-7.9	dp32-14S
1192.174	6	7159.0908	7159.1192	-4.0	dp32-15S
1226.0125	6	7362.1218	7362.1986	-10.4	dp33-15S
1242.0262	6	7458.204	7458.2739	-9.4	dp34-14S
1255.3475	6	7538.1318	7538.2307	-13.1	dp34-15S
1075.87	7	7538.1446	7538.2307	-11.4	dp34-15S
1268.6689	6	7618.0602	7618.1875	-16.7	dp34-16S
1331.8589	6	7997.2002	7997.299	-12.4	dp36-16S
1345.2096	6	8077.3044	8077.2558	6.0	dp36-17S

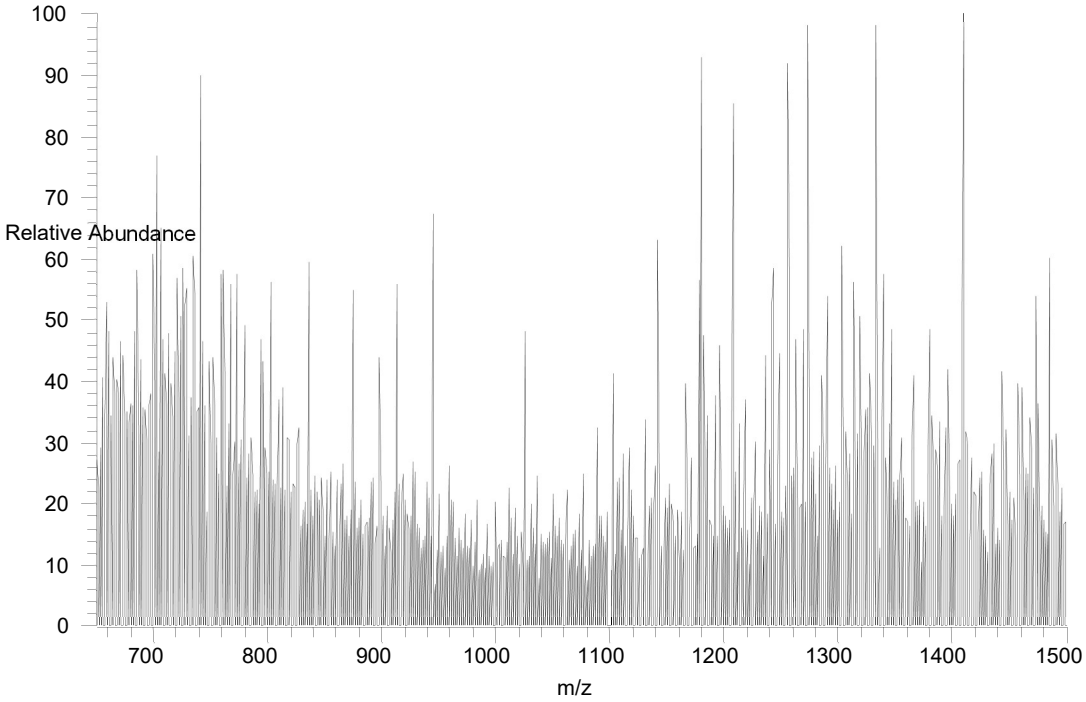
Supplementary Table 4.7. Accurate mass measurements for 17 compositions identified by FTMS of f52 (7.2 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.17. FTMS of fraction 58 (7.9 kDa by PAGE) corresponding to 17 composition identifications listed in Supplementary Table 4.8.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
872.8057	6	5242.881	5242.8892	-1.6	dp24-10S
1047.5747	5	5242.9125	5242.8892	4.4	dp24-10S
1012.5019	6	6081.0582	6081.069	-1.8	dp28-11S
1025.8315	6	6161.0358	6161.0258	1.6	dp28-12S
1072.9952	6	6444.018	6444.062	-6.8	dp29-13S
1089.0139	6	6540.1302	6540.1373	-1.1	dp30-12S
1102.3443	6	6620.1126	6620.0941	2.8	dp30-13S
1115.6668	6	6700.0476	6700.0509	-0.5	dp30-14S
1165.5292	6	6999.222	6999.2056	2.3	dp32-13S
1178.848	6	7079.1348	7079.1624	-3.9	dp32-14S
1242.0364	6	7458.2652	7458.2739	-1.2	dp34-14S
1255.3686	6	7538.2584	7538.2307	3.7	dp34-15S
1302.5358	6	7821.2616	7821.2669	-0.7	dp35-16S
1318.5463	6	7917.3246	7917.3422	-2.2	dp36-15S
1331.8834	6	7997.3472	7997.299	6.0	dp36-16S
1395.0581	6	8376.3954	8376.4105	-1.8	dp38-16S
1408.405	6	8456.4768	8456.3673	12.9	dp38-17S
1207.0535	7	8456.4291	8456.3673	7.3	dp38-17S
1272.6219	7	8915.4079	8915.4356	-3.1	dp40-18S

Supplementary Table 4.8. Accurate mass measurements for 17 compositions identified by FTMS of f58 (7.9 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.



Supplementary Figure 4.18. FTMS of fraction 60 (8.7 kDa by PAGE) corresponding to 29 composition identifications listed in Supplementary Table 4.9.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
1025.8312	6	6161.034	6161.0258	1.3	dp28-12S
1102.3465	6	6620.1258	6620.0941	4.8	dp30-13S
1178.8528	6	7079.1636	7079.1624	0.2	dp32-14S
1226.0306	6	7362.2304	7362.1986	4.3	dp33-15S
1050.7348	7	7362.1982	7362.1986	-0.1	dp33-15S
1242.0395	6	7458.2838	7458.2739	1.3	dp34-14S
1255.3643	6	7538.2326	7538.2307	0.3	dp34-15S
1075.8930	7	7538.3056	7538.2307	9.9	dp34-15S
1268.6938	6	7618.2096	7618.1875	2.9	dp34-16S
1289.2120	6	7741.3188	7741.3101	1.1	dp35-15S
1302.5353	6	7821.2586	7821.2669	-1.1	dp35-16S
1130.0454	7	7917.3724	7917.3422	3.8	dp36-15S
1318.5493	6	7917.3426	7917.3422	0.1	dp36-15S
1331.8881	6	7997.3754	7997.299	9.6	dp36-16S
1141.4683	7	7997.3327	7997.299	4.2	dp36-16S
1345.2100	6	8077.3068	8077.2558	6.3	dp36-17S
744.4833	11	8200.4021	8200.3784	2.9	dp37-16S
1365.7198	6	8200.3656	8200.3784	-1.6	dp37-16S
1181.9020	7	8280.3686	8280.3352	4.0	dp37-17S
1379.0452	6	8280.318	8280.3352	-2.1	dp37-17S
760.4894	11	8376.4692	8376.4105	7.0	dp38-16S
836.6386	10	8376.464	8376.4105	6.4	dp38-16S
1195.6294	7	8376.4604	8376.4105	6.0	dp38-16S

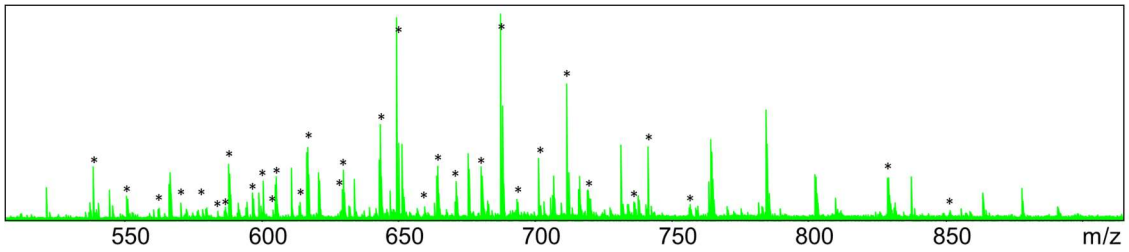
1395.0713	6	8376.4746	8376.4105	7.7	dp38-16S
703.6920	12	8456.3976	8456.3673	3.6	dp38-17S
767.7564	11	8456.4062	8456.3673	4.6	dp38-17S
1207.0514	7	8456.4144	8456.3673	5.6	dp38-17S
1408.3883	6	8456.3766	8456.3673	1.1	dp38-17S
713.9507	12	8579.502	8579.4899	1.4	dp39-16S
1236.0504	7	8659.4074	8659.4467	-4.5	dp39-17S
1442.2332	6	8659.446	8659.4467	-0.1	dp39-17S

Supplementary Table 4.9. Accurate mass measurements for 29 compositions identified by FTMS of f60 (8.7 kDa by PAGE). Data acquired on the Orbitrap FT mass spectrometer.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
727.2800	12	8739.4536	8739.4035	5.7	dp39-18S
793.4847	11	8739.4175	8739.4035	1.6	dp39-18S
1247.4770	7	8739.3936	8739.4035	-1.1	dp39-18S
735.2841	12	8835.5028	8835.4788	2.7	dp40-17S
1261.1999	7	8835.4539	8835.4788	-2.8	dp40-17S
741.9466	12	8915.4528	8915.4356	1.9	dp40-18S
809.4868	11	8915.4406	8915.4356	0.6	dp40-18S
1272.6299	7	8915.4639	8915.4356	3.2	dp40-18S
1484.9055	6	8915.4798	8915.4356	5.0	dp40-18S
1284.0530	7	8995.4256	8995.3924	3.7	dp40-19S
1290.2129	7	9038.5449	9038.5582	-1.5	dp41-17S
758.8705	12	9118.5396	9118.515	2.7	dp41-18S
1301.6429	7	9118.5549	9118.515	4.4	dp41-18S
765.5347	12	9198.51	9198.4718	4.2	dp41-19S
706.5675	13	9198.4789	9198.4718	0.8	dp41-19S
1313.0596	7	9198.4718	9198.4718	0	dp41-19S
1532.0730	6	9198.4848	9198.4718	1.4	dp41-19S
773.5395	12	9294.5676	9294.5471	2.2	dp42-18S
1326.7881	7	9294.5713	9294.5471	2.6	dp42-18S
780.2043	12	9374.5452	9374.5039	4.4	dp42-19S
720.1122	13	9374.56	9374.5039	6.0	dp42-19S
1338.2081	7	9374.5113	9374.5039	0.8	dp42-19S
1561.4119	6	9374.5182	9374.5039	1.5	dp42-19S

1378.6394	7	9657.5304	9657.5401	-1.0	dp43-20S
1403.7905	7	9833.5881	9833.5722	1.6	dp44-20S

Supplementary Table 4.9. (continued)

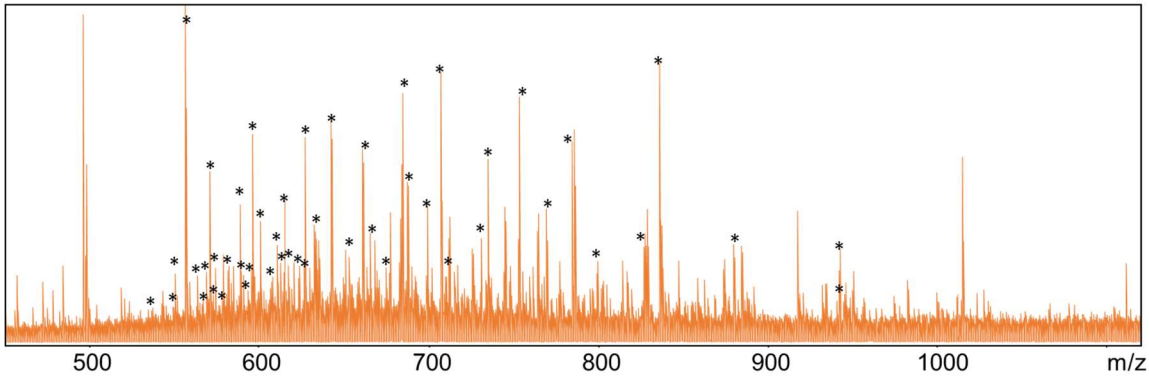


Supplementary Figure 4.19. FT-ICR MS of fraction 35 (4.2 kDa by PAGE) corresponding to 21 composition identifications listed in Supplementary Table 4.10. * - indicates a peak that could be assigned a decorin composition.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
715.8896	4	2867.5896	2867.5909	-0.4	dp14-4S
572.5101	5	2867.5895	2867.5909	-0.4	dp14-4S
954.8582	3	2867.5980	2867.5909	2.6	dp14-4S
588.5012	5	2947.5450	2947.5477	-0.9	dp14-5S
735.8793	4	2947.5484	2947.5477	0.3	dp14-5S
981.5101	3	2947.5537	2947.5477	2.1	dp14-5S
604.4926	5	3027.5020	3027.5045	-0.7	dp14-6S
629.1175	5	3150.6265	3150.6271	-0.1	dp15-5S
830.6568	4	3326.6584	3326.6592	-0.2	dp16-5S
664.3240	5	3326.6590	3326.6592	0.0	dp16-5S
680.3151	5	3406.6145	3406.6160	-0.4	dp16-6S
850.6478	4	3406.6224	3406.6160	2.0	dp16-6S
580.0873	6	3486.5706	3486.5728	-0.6	dp16-7S
600.6085	6	3609.6978	3609.6954	0.8	dp17-6S
613.9336	6	3689.6484	3689.6522	0.9	dp17-7S
740.1458	5	3705.7680	3705.7707	-0.6	dp18-5S
629.9464	6	3785.7252	3785.7275	-0.5	dp18-6S
756.1374	5	3785.7260	3785.7275	-0.4	dp18-6S
643.2726	6	3865.6824	3865.6843	-0.4	dp18-7S
551.2327	7	3865.6835	3865.6843	-0.2	dp18-7S
562.6547	7	3945.6375	3945.6411	-0.9	dp18-8S
591.6656	7	4148.7138	4148.7205	-1.6	dp19-8S
693.1316	6	4164.8364	4164.8390	-0.5	dp20-6S

605.3910	7	4244.7916	4244.7958	-0.9	dp20-7S
706.4578	6	4244.7936	4244.7958	-0.4	dp20-7S
719.7833	6	4324.7466	4324.7526	-1.3	dp20-8S
539.5858	8	4324.7488	4324.7526	-0.8	dp20-8S
616.8137	7	4324.7505	4324.7526	-0.4	dp20-8S
659.5490	7	4623.8976	4623.9073	-2.1	dp22-7S
670.9724	7	4703.8614	4703.8641	-0.6	dp22-8S
586.9753	8	4703.8648	4703.8641	0.2	dp22-8S
596.9697	8	4783.8200	4783.8209	-0.1	dp22-9S
682.3964	7	4783.8294	4783.8209	1.8	dp22-9S

Supplementary Table 4.10. Accurate mass measurements for 20 compositions identified by FT-ICR MS of f35 (4.2 kDa by PAGE). Data acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer



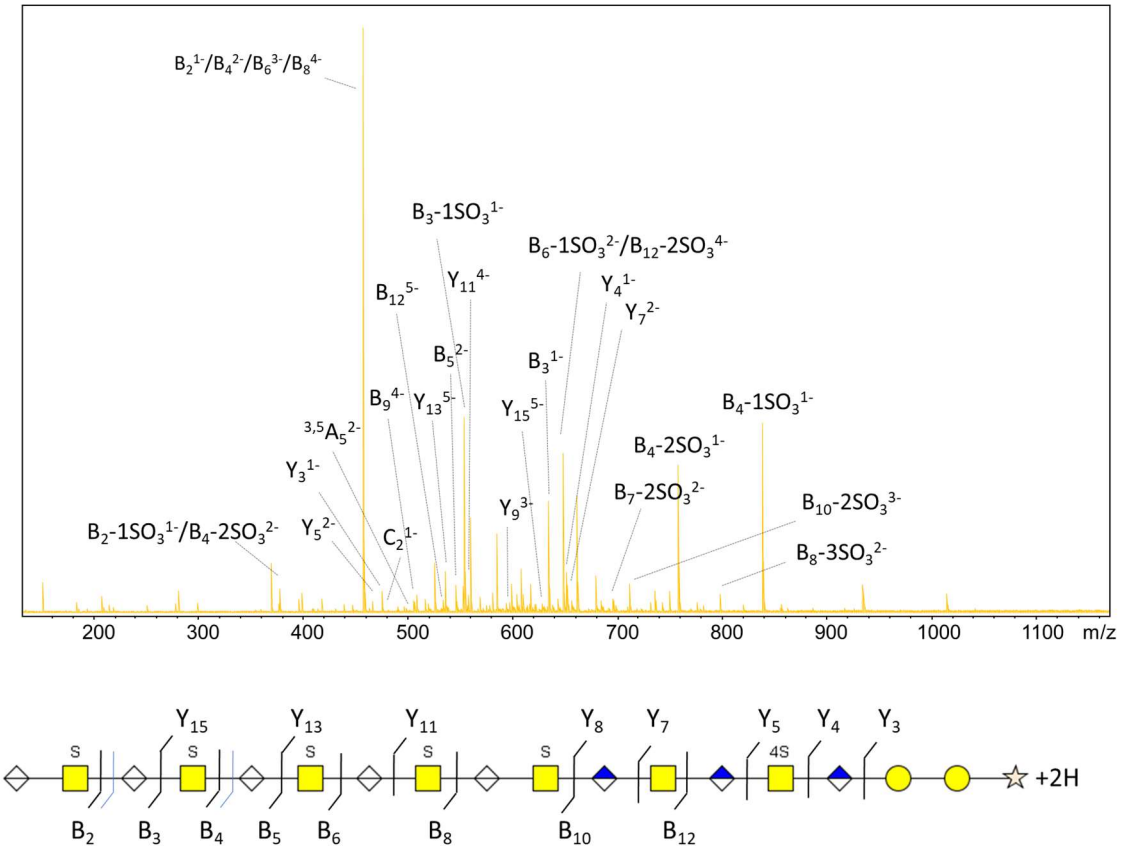
Supplementary Figure 4.20. FT-ICR MS of fraction 51 (7.0 kDa by PAGE) corresponding to 15 composition identifications listed in Supplementary Table 4.11. * - indicates a peak that could be assigned a decorin composition.

m/z	z	M_{Exp}	M_{Theor}	Accuracy(ppm)	Chain
596.9682	8	4783.803	4783.8209	-2.6	dp22-9S
581.5354	9	5242.8884	5242.8892	-0.1	dp24-10S
654.3517	8	5242.876	5242.8892	-2.5	dp24-10S
561.1903	10	5621.9806	5622.0007	-3.5	dp26-10S
623.6581	9	5621.9929	5622.0007	-1.4	dp26-10S
569.1871	10	5701.9486	5701.9575	-1.5	dp26-11S
632.5425	9	5701.9526	5701.9575	-0.8	dp26-11S
711.7353	8	5701.9452	5701.9575	-2.1	dp26-11S
607.0982	10	6081.0603	6081.069	-1.4	dp28-11S
674.6653	9	6081.0579	6081.069	-1.8	dp28-11S
683.5495	9	6161.0158	6081.069	-1.6	dp28-11S
559.082	11	6160.9873	6161.0258	-6.2	dp28-12S
615.0969	10	6161.0469	6161.0258	-1.6	dp28-12S
593.549	11	6540.1252	6540.1373	-1.8	dp30-12S
725.6739	9	6540.1356	6540.1373	-0.2	dp30-12S
826.5018	8	6620.0767	6540.1373	-2.6	dp30-12S
550.6652	12	6620.0761	6620.0941	-2.7	dp30-13S
600.8179	11	6620.0822	6620.0941	-1.8	dp30-13S
661.0011	10	6620.089	6620.0941	-0.7	dp30-13S
734.5546	9	6620.0615	6620.0941	-3.4	dp30-13S
944.7185	7	6620.0838	6620.0941	-1.5	dp30-13S
574.2499	12	6903.092	6903.1303	-5.5	dp31-14S
582.2581	12	6999.1905	6999.2056	-2.1	dp32-13S

635.2814	11	6999.1814	6999.2056	-3.4	dp32-13S
698.9127	10	6999.2048	6999.2056	-0.1	dp32-13S
776.6797	9	6999.1872	6999.2056	-2.6	dp32-13S
543.5408	13	7079.1312	7079.1624	-4.4	dp32-14S
588.9217	12	7079.1535	7079.1624	3.2	dp32-14S
642.5506	11	7079.142	7079.1624	-2.8	dp32-14S
706.9076	10	7079.1535	7079.1624	-1.2	dp32-14S
785.565	9	7079.1553	7079.1624	-1.0	dp32-14S
883.8833	8	7079.1288	7079.1624	-4.7	dp32-14S
595.5842	12	7159.1037	7159.1192	-2.1	dp32-15S
565.3129	13	7362.1692	7362.1986	-4.0	dp33-15S
612.5067	12	7362.1745	7362.1986	-3.2	dp33-15S
667.0166	11	7458.2683	7458.2739	-0.7	dp34-14S
572.7029	13	7458.2396	7458.2739	-4.6	dp34-14S
620.5134	12	7458.2538	7458.2739	-2.7	dp34-14S
537.4364	14	7538.2189	7538.2307	-1.5	dp34-15S
578.8551	13	7538.2174	7538.2307	-1.7	dp34-15S
627.1771	12	7538.219	7538.2307	-1.5	dp34-15S
684.2843	11	7538.2132	7538.2307	-2.3	dp34-15S
752.8086	10	7538.1637	7538.2307	-8.9	dp34-15S
836.5712	9	7538.2114	7538.2307	-2.5	dp34-15S
941.2652	8	7538.184	7538.2307	-6.1	dp34-15S
570.2252	14	7997.2616	7997.299	-4.6	dp36-16S
614.1671	13	7997.2738	7997.299	-3.1	dp36-16S

798.7207	10	7997.2848	7997.299	2.9	dp36-16S

Supplementary Table 4.11. Accurate mass measurements for 15 compositions identified by FT-ICR MS of f51 (7.0 kDa by PAGE). Data acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer



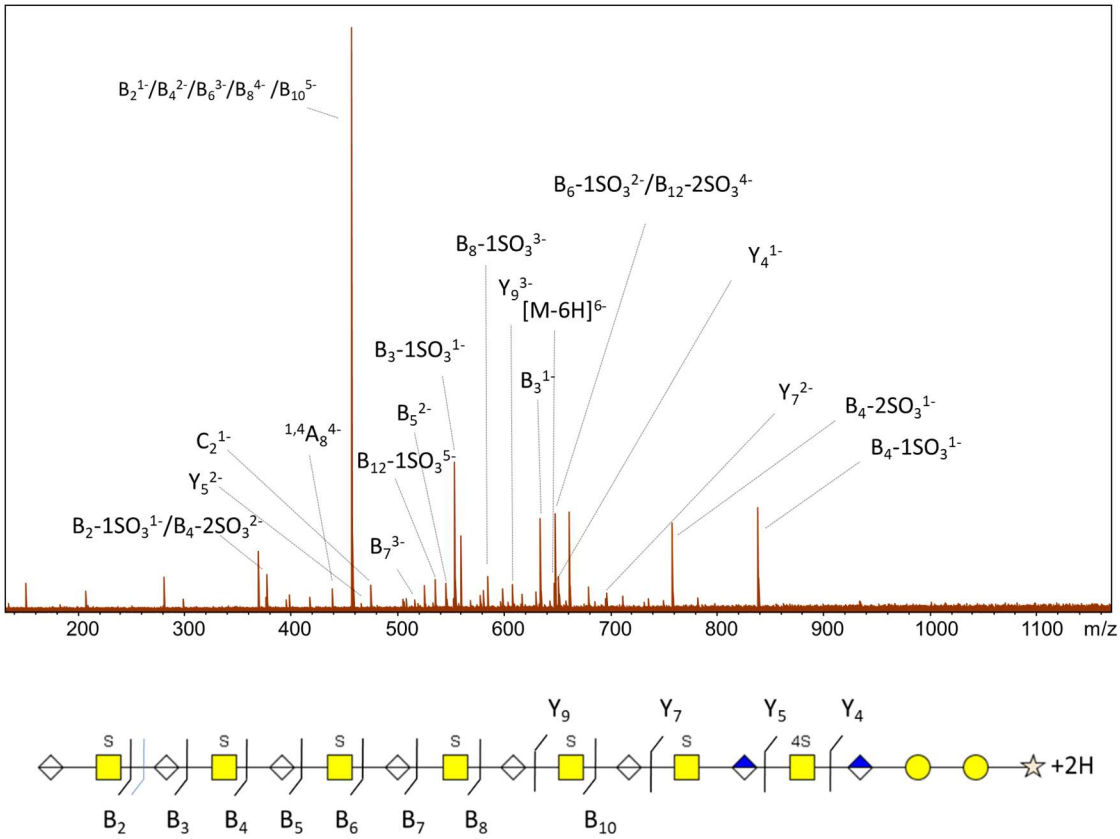
Supplementary Figure 4.21. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 629.9464 (z=6) and its fragmentation pattern providing sequence for composition dp18-6S (f35). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.13.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0615	100.00	HexA1HexNAc1S1-B	-1.17	458.0610	0
B02	1	378.1029	3.85	HexA1HexNAc1-B	3.42	378.1042	1
B03	1	634.0922	19.10	HexA2HexNAc1S1-B	1.45	634.0931	0
B04	2	458.0615	100.00	HexA2HexNAc2S2-B	-1.17	458.0610	0
B04	1	837.1749	32.08	HexA2HexNAc2S1-B	-2.92	837.1725	1
B04	2	418.0817	2.09	HexA2HexNAc2S1-B	2.04	418.0826	1
B04	1	757.2152	25.30	HexA2HexNAc2-B	-0.55	757.2157	2
B04	2	378.1029	3.85	HexA2HexNAc2-B	3.42	378.1042	2
B05	2	546.0763	4.40	HexA3HexNAc2S2-B	1.34	546.0768	0
B05	1	1013.2072	2.79	HexA3HexNAc2S1-B	-2.62	1013.2046	1
B05	2	506.0977	1.84	HexA3HexNAc2S1-B	1.76	506.0986	1
B06	3	458.0615	100.00	HexA3HexNAc3S3-B	-1.17	458.0610	0
B06	2	647.6136	26.91	HexA3HexNAc3S2-B	4.80	647.6167	1
B06	3	378.1029	3.85	HexA3HexNAc3-B	3.42	378.1042	3
B07	2	735.6329	3.35	HexA4HexNAc3S2-B	-0.20	735.6328	1
B07	3	490.0846	0.26	HexA4HexNAc3S2-B	3.00	490.0860	1
B07	2	695.6539	1.98	HexA4HexNAc3S1-B	0.62	695.6544	2
B08	4	458.0615	100.00	HexA4HexNAc4S4-B	-1.17	458.0610	0
B08	2	837.1749	32.08	HexA4HexNAc4S2-B	-2.92	837.1725	2
B08	3	557.7783	2.69	HexA4HexNAc4S2-B	1.68	557.7792	2
B08	4	418.0817	2.09	HexA4HexNAc4S2-B	2.04	418.0826	2
B08	2	797.1946	2.73	HexA4HexNAc4S1-B	-0.63	797.1941	3
B09	4	502.0679	0.17	HexA5HexNAc4S4-B	2.24	502.0686	0
B09	3	643.1081	2.04	HexA5HexNAc4S3-B	1.22	643.1088	1
B09	2	925.1885	0.25	HexA5HexNAc4S2-B	0.07	925.1885	2

B09	3	616.4560	1.23	HexA5HexNAc4S2-B	0.90	616.4566	2
B10	4	552.8379	4.10	HexA5HexNAc5S4-B	1.80	552.8389	1
B10	3	710.8013	4.50	HexA5HexNAc5S3-B	0.89	710.8019	2
B10	4	532.8485	0.39	HexA5HexNAc5S3-B	2.13	532.8497	2
B10	3	684.1488	1.79	HexA5HexNAc5S2-B	1.31	684.1497	3
B12	5	533.8825	1.84	HexA6HexNAc6S5-B	1.49	533.8825	0
B12	4	647.6136	26.91	HexA6HexNAc6S4-B	4.79	647.6167	1
B12	3	837.1749	32.08	HexA6HexNAc6S3-B	-2.93	837.1725	2
B12	4	627.6272	0.62	HexA6HexNAc6S3-B	0.54	627.6250	2
B12	3	810.5206	0.14	HexA6HexNAc6S2-B	-0.44	810.5202	3
Y03	1	475.1659	0.30	Y-Hex2Pen1-2H	1.96	475.1668	0
Y04	1	651.1987	6.90	Y-Hex2HexA1Pen1-2H	0.31	651.1989	0
Y05	2	466.6129	1.76	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.97	466.6136	0
Y07	2	656.1689	1.75	Y-Hex2HexA2HexNAc2Pen1S1-2H	1.13	656.1694	0
Y09	3	590.1328	0.41	Y-Hex2HexA3HexNAc3Pen1S2-2H	1.08	590.1331	0
Y11	4	557.1174	0.89	Y-Hex2HexA4HexNAc4Pen1S3-2H	-3.76	577.1149	0
Y13	5	537.3033	0.16	Y-Hex2HexA5HexNAc5Pen1S4-2H	2.10	537.3040	0
Y15	5	629.1167	0.67	Y-Hex2HexA6HexNAc6Pen1S5-2H	2.18	629.1177	0
Y15	5	613.1261	0.63	Y-Hex2HexA6HexNAc6Pen1S4-2H	1.11	613.1267	1
C02	1	476.0705	3.41	HexA1HexNAc1S1-C	2.22	476.0705	0
C02	1	396.1140	2.06	HexA1HexNAc1-C	1.99	396.1147	1
C04	2	467.0651	0.48	HexA2HexNAc2S2-C	2.56	467.0660	0
C04	1	855.1826	1.01	HexA2HexNAc2S1-C	0.54	855.1830	1

C04	1	775.2263	1.39	HexA2HexNAc2S0-C	-0.10	775.2262	2
3,5-A03	1	546.0763	4.40	HexA2HexNAc1S1-A	1.43	546.0771	0
3,5-A05	2	502.0679	0.17	HexA2HexNAc1S1-A	2.29	502.0690	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	16.20	HexA2HexNAc3S2	1.25	559.6007	
	2	370.0466	0.80	HexA1GalNAc2S2	-4.32	370.0450	
	1	300.0389	1.70	GalNAc1S1-O	2.00	300.0395	
	1	282.0283	3.80	GalNAc1S1	2.13	282.0289	
	1	661.1397	20.00	HexA1GalNAc2S1	1.06	661.1404	

Supplementary Table 4.12. Assignment of fragment ions resulted from CID of parent ion m/z 629.9464 (z=6) corresponding to composition dp18-6S (f35). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.

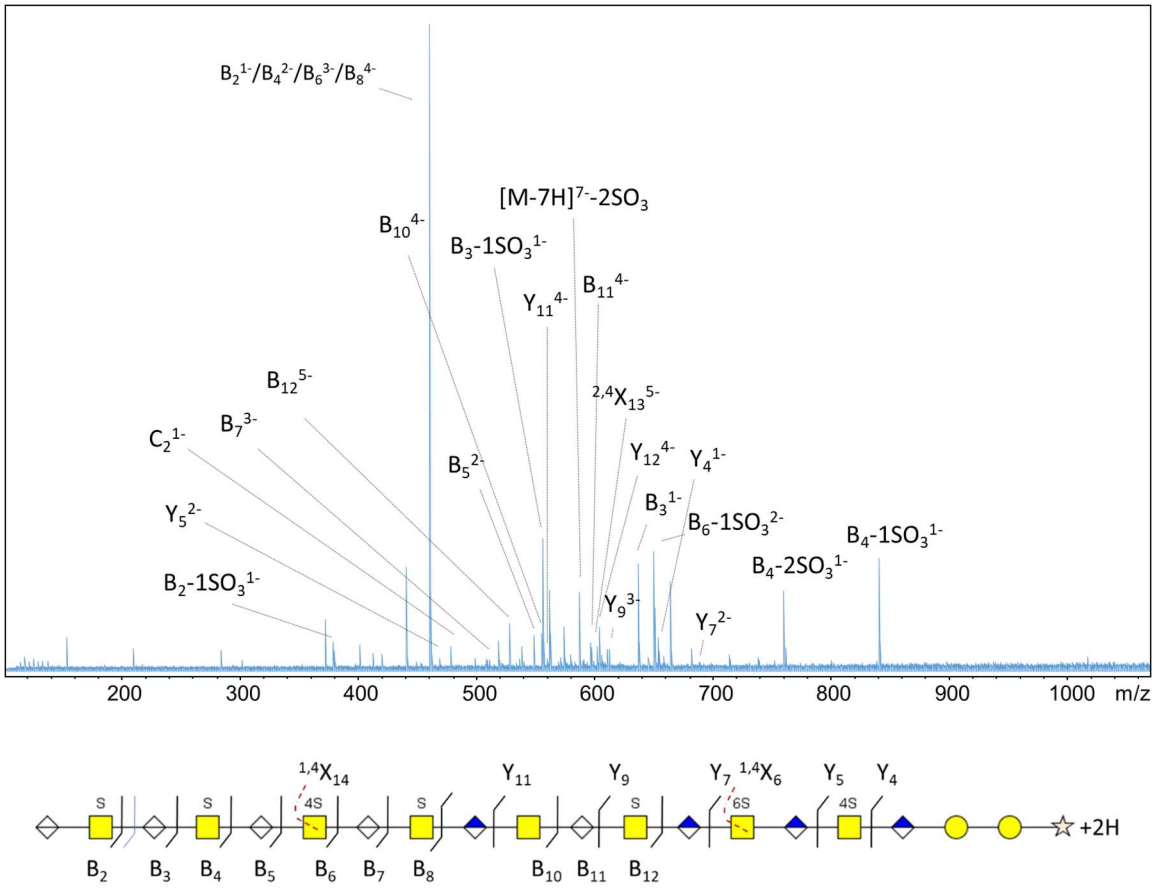


Supplementary Figure 4.22. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 643.2726 (z=6) and its fragmentation pattern providing sequence for composition dp18-7S (f35). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.13.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z THEOR	#SO3 LOSS
B02	1	458.0613	100.00	HexA1HexNAc1S1-B	-0.66	458.0610	0
B03	1	634.0923	15.90	HexA2HexNAc1S1-B	1.29	634.0931	0
B03	1	554.1356	25.40	HexA2HexNAc1S0-B	1.23	554.1363	1
B04	2	458.0613	100.00	HexA2HexNAc2S2-B	-0.66	458.0610	0
B04	2	835.1735	17.90	HexA2HexNAc2S1-B	-1.30	835.1724	1
B04	2	757.2155	15.70	HexA2HexNAc2S0-B	0.26	757.2157	2
B05	2	546.0765	4.09	HexA3HexNAc2S2-B	1.07	546.0770	0
B06	3	458.0613	100.00	HexA3HexNAc3S3-B	-0.66	458.0610	0
B06	2	647.6165	15.85	HexA3HexNAc3S2-B	0.42	647.6167	1
B06	2	607.6375	3.80	HexA3HexNAc3S1-B	1.37	607.6383	2
B07	3	516.7372	1.21	HexA4HexNAc3S3-B	2.16	516.7383	0
B07	2	695.6538	1.25	HexA4HexNAc3S1-B	0.77	695.6544	2
B08	4	458.0613	100.00	HexA4HexNAc4S4-B	-0.66	458.0610	0
B08	3	584.4311	5.08	HexA4HexNAc4S3-B	0.69	584.4315	1
B08	3	557.7785	0.54	HexA4HexNAc4S2-B	1.25	557.7792	2
B08	2	797.1932	0.43	HexA4HexNAc4S1-B	1.13	797.1941	3
B10	5	458.0613	100.00	HexA5HexNAc5S5-B	-0.66	458.0610	0
B10	4	552.8378	0.95	HexA5HexNAc5S4-B	1.94	552.8389	1
B10	3	710.8013	1.73	HexA5HexNAc5S3-B	1.01	710.8020	2
B12	5	533.8825	15.30	HexA6HexNAc6S5-B	1.43	533.8833	1
B12	4	647.6165	15.85	HexA6HexNAc6S4-B	0.42	647.6167	2
Y04	1	651.1988	6.00	Y-Hex2HexA1Pen1-2H	0.18	651.1989	0
Y05	2	466.6132	0.69	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.50	466.6138	0
Y07	2	696.1485	2.17	Y-Hex2HexA2HexNAc2Pen1S1-2H	-0.76	696.1480	0

Y09	3	616.7878	2.06	Y-Hex2HexA3HexNAc3Pen1S2-2H	-3.52	616.7857	0
C02	1	476.0705	3.83	HexA1HexNAc1S1-C	2.16	476.0715	0
3,5-A03	1	546.0765	4.09	HexA2HexNAc1S1-A	1.16	546.0771	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6002	12.90	HexA2GalNAc3S2	0.90	559.6007	
	2	370.0442	10.30	HexA1GalNAc2S2	1.99	370.0450	
	1	300.0389	2.20	GalNAc1S1-O	2.02	300.0395	
	1	282.0283	6.00	GalNAc1S1	2.02	282.0289	
	1	661.1399	17.20	HexA1GalNAc2S1	0.76	661.1404	

Supplementary Table 4.13. Assignment of fragment ions resulted from CID of parent ion m/z 643.2726 (z=6) corresponding to composition dp18-7S (f35). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.

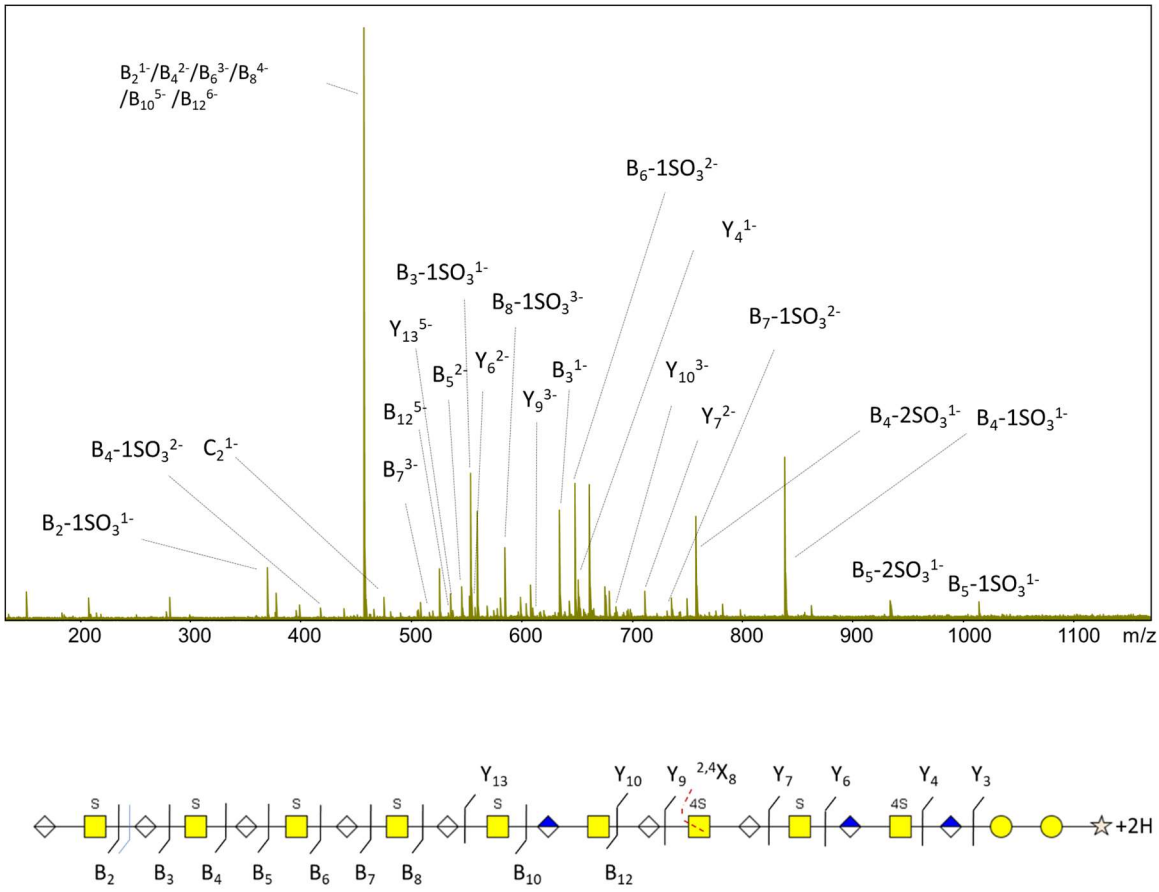


Supplementary Figure 4.23. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 605.3910 (z=7) and its fragmentation pattern providing sequence for composition dp20-7S (f35). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. Diagnostic X_n fragments are shown in red. All fragment ions assigned are provided in Supplementary Table 4.14.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z THEOR	#SO3 LOSS
B02	1	458.0591	100.00	HexA1HexNAc1S1-B	4.08	458.0610	0
B02	1	378.1034	2.39	HexA1HexNAc1-B	2.20	378.1042	1
B03	1	634.0923	15.63	HexA2HexNAc1S1-B	1.30	634.0931	0
B03	1	554.1356	19.50	HexA2HexNAc1-B	1.28	554.1363	1
B04	2	458.0591	100.00	HexA2HexNAc2S2-B	4.08	458.0610	0
B04	2	418.0815	1.84	HexA2HexNAc2S1-B	2.52	418.0826	1
B04	1	757.2154	11.30	HexA2HexNAc2-B	0.31	757.2157	2
B04	2	378.1034	2.39	HexA2HexNAc2-B	2.20	378.1042	2
B05	2	546.0762	4.60	HexA3HexNAc2S2-B	1.47	546.0770	0
B06	3	458.0591	100.00	HexA3HexNAc3S3-B	4.08	458.0610	0
B06	2	647.6166	19.90	HexA3HexNAc3S2-B	0.15	647.6167	1
B06	2	607.6377	2.38	HexA3HexNAc3S1-B	1.06	607.6383	2
B06	3	378.1034	2.39	HexA3HexNAc3-B	2.20	378.1042	3
B07	3	516.7374	3.75	HexA4HexNAc3S3-B	1.78	516.7383	0
B07	2	735.6322	1.07	HexA4HexNAc3S2-B	0.72	735.6328	1
B07	2	695.6548	0.03	HexA4HexNAc3S1-B	-0.60	695.6544	2
B08	4	458.0591	100.00	HexA4HexNAc4S4-B	4.08	458.0610	0
B08	3	557.7775	0.01	HexA4HexNAc4S2-B	3.12	557.7792	2
B08	4	418.0815	1.84	HexA4HexNAc4S2-B	2.52	418.0826	2
B09	3	643.1085	1.12	HexA5HexNAc4S3-B	0.58	643.1088	1
B10	4	552.8378	4.88	HexA5HexNAc5S4-B	1.99	552.8388	0
B10	3	710.8010	1.57	HexA5HexNAc5S3-B	1.37	710.8020	1
B11	4	596.8462	0.19	HexA6HexNAc5S4-B	1.16	596.8469	0
B12	5	533.8827	1.06	HexA6HexNAc6S5-B	1.10	533.8833	0
B12	6	418.0815	1.84	HexA6HexNAc6S3-B	2.51	418.0826	2

B12	4	607.6377	2.38	HexA6HexNAc6S2-B	1.05	607.6383	3
B14	5	609.7042	1.47	HexA7HexNAc7S5-B	2.20	609.7056	1
Y04	1	651.1986	5.30	Y-Hex2HexA1Pen1-2H	0.43	651.1989	0
Y05	2	466.6131	1.22	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.52	466.6138	0
Y07	2	696.1487	0.31	Y-Hex2HexA2HexNAc2Pen1S2-2H	-1.05	696.1480	0
Y09	3	616.7857	0.17	Y-Hex2HexA3HexNAc3Pen1S3-2H	-0.09	616.7857	0
Y11	4	557.1151	1.02	Y-Hex2HexA4HexNAc4Pen1S3-2H	0.41	557.1153	0
Y12	4	601.1218	5.87	Y-Hex2HexA5HexNAc4Pen1S3-2H	2.48	601.1233	0
C02	1	476.0704	2.95	HexA1HexNAc1S1-C	2.51	476.0715	0
3,5-A03	1	546.0762	4.60	HexA2HexNAc1S1-A	1.56	546.0771	0
1,4-X06	2	630.6201	0.16	X-Hex2HexA2HexNAc1Pen1S1-2H	-2.01	630.6188	0
2,4-X14	5	601.1218	5.87	X-Hex2HexA6HexNAc5Pen1S3-2H	1.11	601.1225	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	12.50	HexA2GalNAc3S2	1.3	559.6007	
	2	370.0442	8.00	HexA1GalNAc2S2	2.2	370.0450	
	1	300.0390	1.80	GalNAc1S1-O	1.7	300.0395	
	1	282.0283	3.30	GalNAc1S1	2.1	282.0289	
	1	661.1398	13.90	HexA1GalNAc2S1	0.9	661.1404	

Supplementary Table 4.14. Assignment of fragment ions resulted from CID of parent ion m/z 605.3910 (z=7) corresponding to composition dp20-7S (f35). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



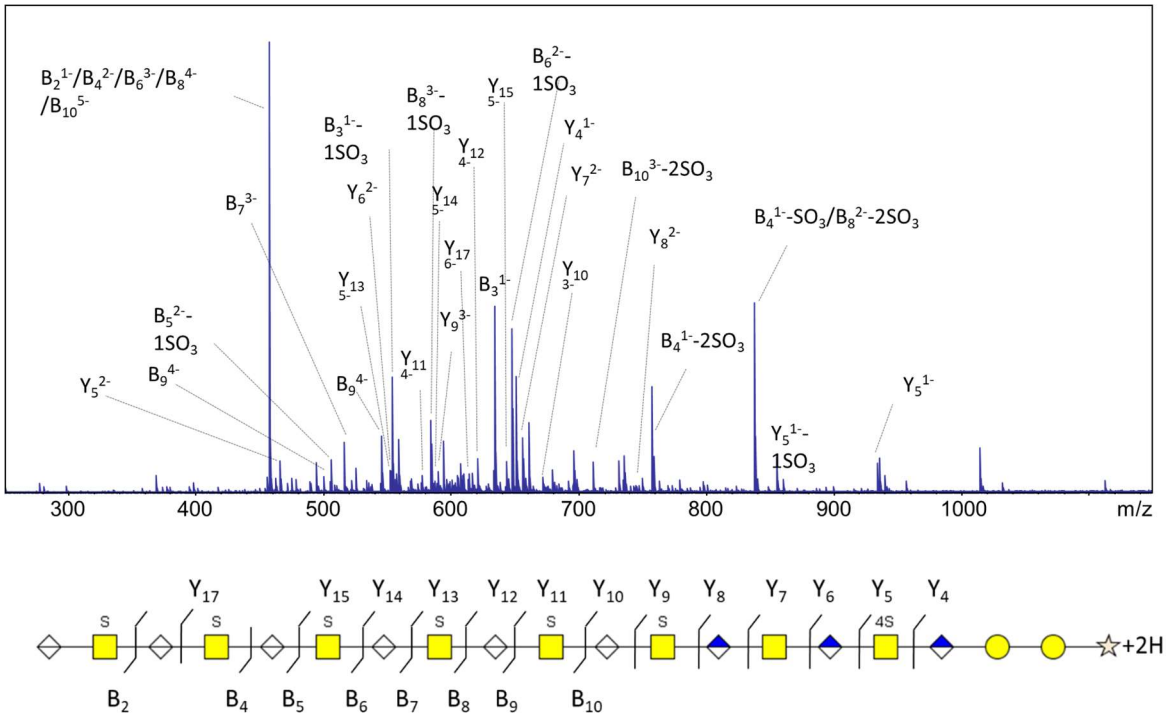
Supplementary Figure 4.24. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 670.9723 (z=7) and its fragmentation pattern providing sequence for composition dp22-8S (f35). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. Diagnostic X_n fragments are shown in red. All fragment ions assigned are provided in Supplementary Table 4.15.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0622	99.49	HexA1HexNAc1S1-B	-2.59	**458.0610**	0
B02	1	378.1033	4.09	HexA1HexNAc1-B	2.41	**378.1042**	1
B03	1	634.0922	18.70	HexA2HexNAc1S1-B	1.46	**634.0931**	0
B03	1	554.1355	24.90	HexA2HexNAc1-B	1.41	**554.1363**	1
B04	2	458.0622	99.49	HexA2HexNAc2S2-B	-2.59	**458.0610**	0
B04	1	837.1733	27.60	HexA2HexNAc2S1-B	-1.04	**837.1724**	1
B04	2	418.0816	1.58	HexA2HexNAc2S1-B	2.30	**418.0826**	1
B04	1	757.2178	16.89	HexA2HexNAc2-B	-2.77	**757.2157**	2
B04	2	378.1033	4.09	HexA2HexNAc2-B	2.41	**378.1042**	2
B05	2	546.0762	5.05	HexA3HexNAc2S2-B	1.62	**546.0770**	0
B05	1	1013.2071	2.25	HexA3HexNAc1S1-B	-2.50	**1013.2046**	1
B05	1	933.2491	2.58	HexA3HexNAc1-B	-1.45	**933.2478**	2
B06	3	458.0622	99.49	HexA3HexNAc3S3-B	-2.59	**458.0610**	0
B06	2	647.6163	23.10	HexA3HexNAc3S2-B	0.71	**647.6168**	1
B06	2	607.6375	5.31	HexA3HexNAc3S1-B	1.40	**607.6383**	2
B06	3	378.1033	4.09	HexA3HexNAc3-B	2.41	**378.1042**	3
B07	3	516.7372	0.82	HexA4HexNAc3S3-B	2.16	**516.7383**	0
B07	2	735.6326	3.07	HexA4HexNAc3S2-B	0.25	**735.6328**	1
B07	2	695.6535	0.85	HexA4HexNAc3S1-B	1.24	**695.6544**	2
B08	4	458.0622	99.49	HexA4HexNAc4S4-B	-2.59	**458.0610**	0
B08	3	584.4310	11.61	HexA4HexNAc4S3-B	0.85	**584.4315**	1
B08	3	557.7783	1.62	HexA4HexNAc4S2-B	1.56	**557.7792**	2
B08	4	418.0816	1.58	HexA4HexNAc4S2-B	2.30	**418.0826**	2
B08	2	797.1946	0.97	HexA4HexNAc4S1-B	-0.66	**797.1941**	3
B10	5	458.0622	99.49	HexA5HexNAc5S5-B	-2.59	**458.0610**	0

B10	4	552.8378	3.47	HexA5HexNAc5S4-B	1.83	**552.8389**	1
B10	3	710.8014	4.17	HexA5HexNAc5S3-B	0.76	**710.8020**	2
B10	3	684.1491	0.34	HexA5HexNAc5S2-B	0.94	**684.1497**	3
B12	5	533.8822	0.64	HexA6HexNAc6S5-B	1.99	**533.8833**	0
B12	4	627.6275	0.05	HexA6HexNAc6S3-B	0.08	**627.6275**	2
B12	6	418.0816	1.58	HexA6HexNAc6S3-B	2.29	**418.0826**	2
B12	4	607.6375	5.31	HexA6HexNAc6S2-B	1.40	**607.6383**	3
B16	6	584.4310	11.61	HexA8HexNAc8S6-B	0.84	**584.4315**	1
B16	6	557.7783	1.62	HexA8HexNAc8S4-B	1.55	**557.7792**	3
Y03	1	475.1662	0.55	Y-Hex2Pen1-2H	1.33	**475.1668**	0
Y04	1	651.1986	6.80	Y-Hex2HexA1Pen1-2H	0.40	**651.1989**	0
Y06	2	554.6290	0.08	Y-Hex2HexA2HexNAc1Pen1S1-2H	1.64	**554.6299**	0
Y07	2	696.1474	1.21	Y-Hex2HexA2HexNAc2Pen1S2-2H	0.83	**696.1480**	0
Y07	2	656.1691	1.19	Y-Hex2HexA2HexNAc2Pen1S1-2H	0.84	**656.1696**	1
Y09	3	616.7885	0.47	Y-Hex2HexA3HexNAc3Pen1S3-2H	-4.63	**616.7857**	0
Y09	3	590.1317	0.18	Y-Hex2HexA3HexNAc3Pen1S2-2H	2.90	**590.1334**	1
Y10	3	675.4621	4.99	Y-Hex2HexA4HexNAc3Pen1S3-2H	1.46	**675.4630**	0
Y10	3	648.8100	0.28	Y-Hex2HexA4HexNAc3Pen1S2-2H	1.21	**648.8108**	1
Y13	5	537.3030	0.21	Y-Hex2HexA5HexNAc5Pen1S4-2H	2.72	**537.3044**	0
C02	1	476.0706	3.37	HexA1HexNAc1S1-C	2.01	**476.0715**	0
C04	1	855.1835	0.60	HexA2HexNAc2S1-C	-0.53	**855.1830**	1
3,5-A03	1	546.0762	5.05	HexA2HexNAc1S1-A	1.71	**546.0771**	0

2,4-X08	2	855.6927	0.08	X-Hex2HexA3HexNAc2Pen1S2-2H	0.54	**855.6931**	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	18.40	HexA2GalNAc3S2	1.30	559.6007	
	2	370.0442	8.85	HexA1GalNAc2S2	1.90	370.0450	
	1	300.0388	0.93	GalNAc1S1-O	2.15	300.0395	
	1	282.0283	3.87	GalNAc1S1	2.24	282.0289	
	1	661.1397	22.00	HexA1GalNAc2S1	1.07	661.1404	

Supplementary Table 4.15. Assignment of fragment ions resulted from CID of parent ion m/z 670.9723 (z=7) corresponding to composition dp22-8S (f35). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



Supplementary Figure 4.25. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 605.3913 (z=7) and its fragmentation pattern providing sequence for composition dp20-7S (f39). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.16.

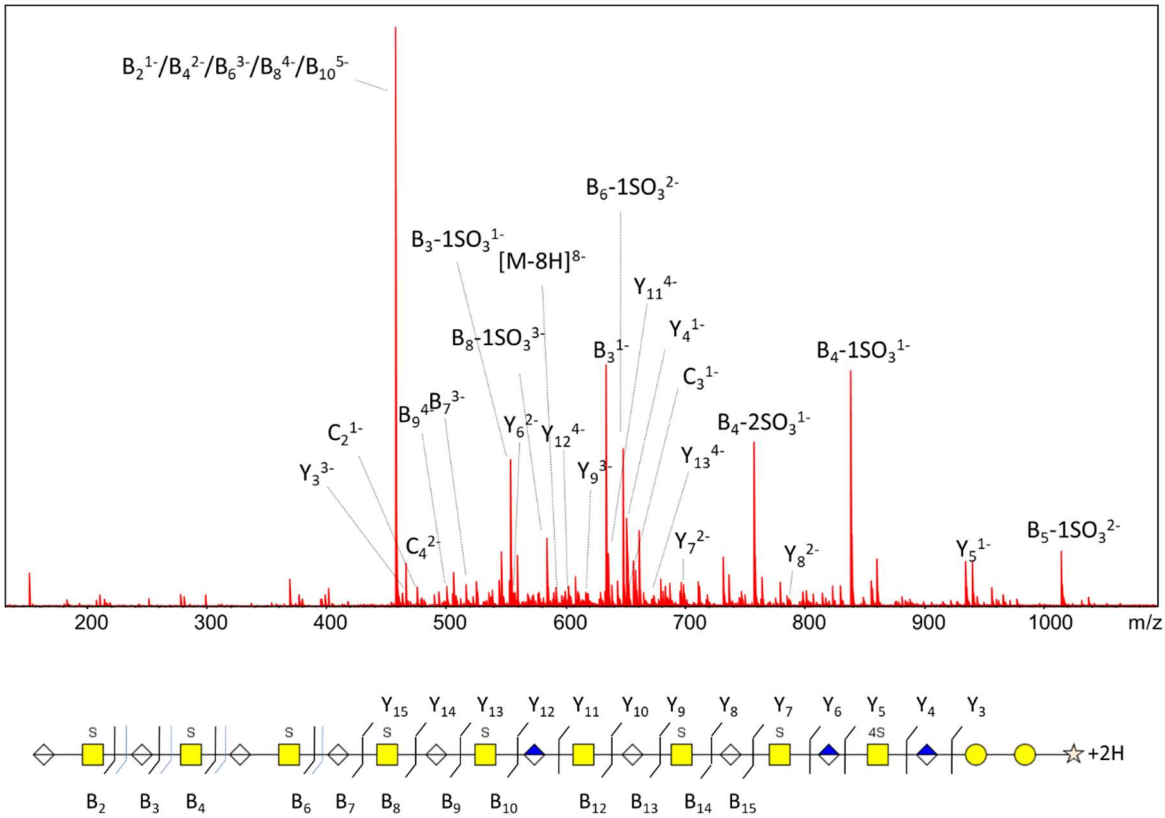
CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0628	99.70	HexA1HexNAc1S1-B	-4.03	458.0610	0
B03	1	554.1350	25.65	HexA2HexNAc1-B	2.29	554.1363	1
B04	2	458.0628	99.70	HexA2HexNAc2S2-B	-4.03	458.0610	0
B04	1	837.1731	41.71	HexA2HexNAc2S1-B	-0.81	837.1725	1
B04	2	418.0820	0.95	HexA2HexNAc2S1-B	1.34	418.0826	1
B04	1	757.2172	23.10	HexA2HexNAc2-B	-2.06	757.2157	2
B05	2	546.0766	12.60	HexA3HexNAc2S2-B	0.77	546.0770	0
B05	2	506.0979	2.43	HexA3HexNAc2S1-B	1.37	506.0986	1
B06	3	458.0628	99.70	HexA3HexNAc3S3-B	-4.03	458.0610	0
B06	2	647.6162	36.50	HexA3HexNAc3S2-B	8.01	647.6214	1
B06	3	431.4078	0.05	HexA3HexNAc3S2-B	2.10	431.4087	1
B06	2	607.6378	5.99	HexA3HexNAc3S1-B	0.93	607.6383	2
B07	3	516.7367	11.20	HexA4HexNAc3S3-B	3.25	516.7384	0
B08	4	458.0628	99.70	HexA4HexNAc4S4-B	-4.03	458.0610	0
B08	3	584.4297	15.76	HexA4HexNAc4S3-B	2.99	584.4315	1
B08	2	837.1731	41.71	HexA4HexNAc4S2-B	-0.81	837.1725	2
B08	4	418.0820	0.95	HexA4HexNAc4S2-B	1.34	418.0826	2
B08	2	797.1946	2.05	HexA4HexNAc4S1-B	-0.71	797.1941	3
B09	4	502.0681	0.36	HexA5HexNAc4S4-B	1.73	502.0690	0
B09	2	925.1917	0.40	HexA5HexNAc4S2-B	-3.41	925.1885	2
B09	3	616.4564	1.04	HexA5HexNAc4S2-B	0.32	616.4566	2
B10	5	458.0628	99.70	HexA5HexNAc5S5-B	-4.03	458.0610	0
B10	4	552.8385	4.62	HexA5HexNAc5S4-B	0.69	552.8389	1
B10	3	710.8022	6.90	HexA5HexNAc5S3-B	-0.32	710.8020	2
B10	4	532.8489	0.59	HexA5HexNAc5S3-B	1.49	532.8497	2

B12	6	458.0628	99.70	HexA6HexNAc6S5-B	-4.03	458.0610	1
B16	6	584.4297	15.76	HexA8HexNAc8S5-B	2.99	584.4315	2
B16	5	685.5277	0.60	HexA8HexNAc8S4-B	0.32	685.5279	3
Y04	1	651.1985	25.70	Y-Hex2HexA1Pen1-2H	0.65	651.1989	0
Y05	1	934.2315	7.32	Y-Hex2HexA1HexNAc1Pen1S1-2H	3.75	934.2350	0
Y05	2	466.6132	6.76	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.41	466.6138	0
Y05	1	854.2750	7.65	Y-Hex2HexA1HexNAc1Pen1-2H	3.75	854.2782	1
Y06	2	554.6293	3.40	Y-Hex2HexA2HexNAc1Pen1S1-2H	1.10	554.6299	0
Y06	1	1030.3134	1.74	Y-Hex2HexA2HexNAc1Pen1-2H	-2.96	1030.3103	1
Y07	2	656.1706	11.86	Y-Hex2HexA2HexNAc2Pen1S1-2H	-1.45	656.1696	0
Y08	2	744.1851	0.38	Y-Hex2HexA3HexNAc2Pen1S1-2H	0.79	744.1856	0
Y09	3	590.1329	4.35	Y-Hex2HexA3HexNAc3Pen1S2-2H	0.78	590.1334	0
Y10	3	675.4631	0.96	Y-Hex2HexA4HexNAc3Pen1S2-2H	-0.05	675.4630	0
Y11	4	577.1063	1.86	Y-Hex2HexA4HexNAc4Pen1S3-2H	-3.15	577.1045	0
Y11	3	743.1560	0.80	Y-Hex2HexA4HexNAc4Pen1S2-2H	0.29	743.1562	1
Y11	4	557.1146	2.68	Y-Hex2HexA4HexNAc4Pen1S2-2H	1.31	557.1153	1
Y11	3	716.5044	0.70	Y-Hex2HexA4HexNAc4Pen1S1-2H	-0.75	716.5039	2
Y12	4	621.1119	3.22	Y-Hex2HexA5HexNAc4Pen1S3-2H	1.09	621.1125	0
Y13	5	553.2948	0.42	Y-Hex2HexA5HexNAc5Pen1S4-2H	1.81	553.2958	0

Y13	4	671.8854	2.49	Y-Hex2HexA5HexNAc5Pen1S3-2H	-4.54	671.8824	1
Y13	5	537.3037	0.21	Y-Hex2HexA5HexNAc5Pen1S3-2H	1.43	537.3044	1
Y14	5	588.5017	0.35	Y-Hex2HexA6HexNAc5Pen1S4-2H	0.92	588.5022	0
Y14	5	572.5102	0.36	Y-Hex2HexA6HexNAc5Pen1S3-2H	1.24	572.5109	1
Y14	4	695.8997	0.23	Y-Hex2HexA6HexNAc5Pen1S2-2H	2.23	695.9012	2
Y15	5	645.1125	2.76	Y-Hex2HexA6HexNAc6Pen1S5-2H	-4.76	645.1095	0
Y15	5	613.1263	2.61	Y-Hex2HexA6HexNAc6Pen1S3-2H	0.77	613.1267	2
Y17	6	613.9364	1.02	Y-Hex2HexA7HexNAc7Pen1S6-2H	-2.74	613.9347	0
Y17	5	720.9350	0.15	Y-Hex2HexA7HexNAc7Pen1S5-2H	-4.45	720.9318	1
Y17	6	600.6084	0.80	Y-Hex2HexA7HexNAc7Pen1S5-2H	0.24	600.6086	1
C04	1	855.1839	2.34	HexA2HexNAc2S1-C	-1.02	855.1830	1
C04	1	775.2265	0.64	HexA2HexNAc2-C	-0.32	775.2262	2
C06	2	656.6215	1.71	HexA3HexNAc3S2-C	0.79	656.6220	1
1,4-A03	1	548.0923	0.51	HexA2HexNAc1S1-A	0.79	548.0927	0
1,5-A08	3	601.7540	0.35	HexA2HexNAc1S4-A	-3.20	601.7521	0
3,5-A05	2	502.0681	0.36	HexA2HexNAc1S2-A	1.78	502.0690	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	11.96	HexA2GalNAc3S2	1.19	559.6007	
	2	370.0443	3.93	HexA1GalNAc2S2	1.87	370.0450	
	1	300.0390	1.52	GalNAc1S1-O	1.52	300.0395	
	1	282.0284	1.36	GalNAc1S1	1.88	282.0289	

1	661.1399	15.74	HexA1GalNAc2S1	0.67	661.1404

Supplementary Table 4.16. Assignment of fragment ions resulted from CID of parent ion m/z 605.3913 (z=7) corresponding to composition dp20-7S (f39). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



Supplementary Figure 4.26. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 586.9750 (z=8) and its fragmentation pattern providing sequence for composition dp22-8S (f39). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.17.

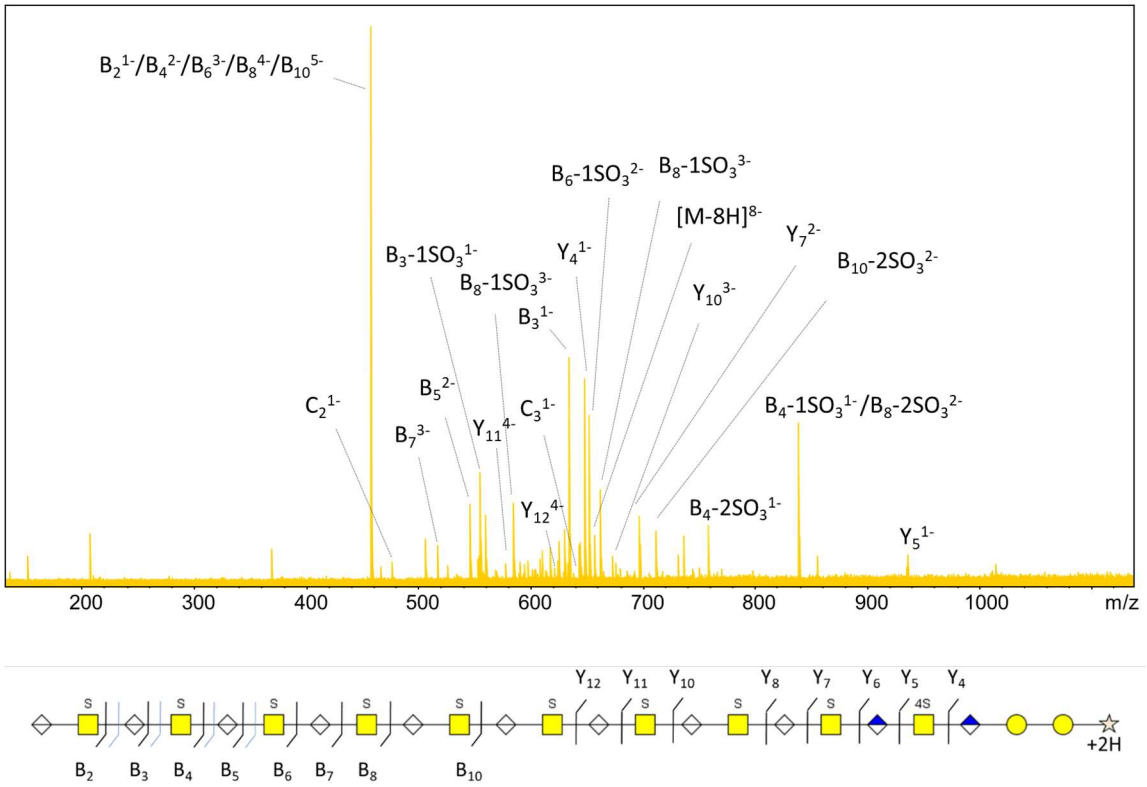
CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0604	100.00	HexA1HexNAc1S1-B	1.21	458.0610	0
B02	1	378.10346	2.00	HexA1HexNAc1-B	1.91	378.1042	1
B03	1	634.09282	41.85	HexA2HexNAc1S1-B	0.41	634.0931	0
B03	1	554.1356	25.60	HexA2HexNAc1-B	1.25	554.1363	1
B04	2	458.0604	100.00	HexA2HexNAc2S2-B	1.21	458.0610	0
B04	1	837.17294	40.88	HexA2HexNAc2S1-B	-0.57	837.1725	1
B04	2	418.08193	0.90	HexA2HexNAc2S1-B	1.56	418.0826	1
B04	1	757.21596	28.49	HexA2HexNAc2-B	-0.39	757.2157	2
B04	2	378.10346	2.00	HexA2HexNAc2-B	1.91	378.1042	2
B05	1	1013.207	9.64	HexA3HexNAc2S1-B	-2.37	1013.2046	1
B05	2	506.09782	2.17	HexA3HexNAc2S1-B	1.60	506.0986	1
B05	1	933.24946	7.79	HexA3HexNAc2-B	-1.82	933.2478	2
B06	3	458.0604	100.00	HexA3HexNAc3S3-B	1.21	458.0610	0
B06	2	647.61613	27.30	HexA3HexNAc3S2-B	0.92	647.6167	1
B06	2	607.63777	5.20	HexA3HexNAc3S1-B	0.91	607.6383	2
B06	3	378.10346	2.00	HexA3HexNAc3-B	1.91	378.1042	3
B07	3	516.73773	3.79	HexA4HexNAc3S3-B	1.20	516.7383	0
B07	2	735.63264	5.51	HexA4HexNAc3S2-B	0.18	735.6328	1
B07	3	490.08584	0.42	HexA4HexNAc3S2-B	0.49	490.0861	1
B07	2	695.65442	2.66	HexA4HexNAc3S1-B	-0.07	695.6544	2
B07	2	655.67413	0.23	HexA4HexNAc3-B	2.81	655.6760	3
B08	4	458.0604	100.00	HexA4HexNAc4S4-B	1.21	458.0610	0
B08	3	584.43094	11.87	HexA4HexNAc4S3-B	0.92	584.4315	1
B08	4	438.07124	0.23	HexA4HexNAc4S3-B	1.24	438.0718	1
B08	2	837.17294	40.88	HexA4HexNAc4S2-B	-0.57	837.1725	2

B08	3	557.7 7862	2.47	HexA4HexNAc4S2 -B	1.06	557.779 2	2
B08	4	418.0 8193	0.90	HexA4HexNAc4S2 -B	1.56	418.082 6	2
B08	2	797.1 9413	2.54	HexA4HexNAc4S1 -B	-0.08	797.194 1	3
B09	4	502.0 6842	1.03	HexA5HexNAc4S4 -B	1.17	502.069 0	0
B09	3	643.1 0838	4.41	HexA5HexNAc4S3 -B	0.72	643.108 8	1
B09	2	925.1 9101	0.77	HexA5HexNAc4S2 -B	-2.70	925.188 5	2
B09	3	616.4 5594	1.08	HexA5HexNAc4S2 -B	1.03	616.456 6	2
B09	2	885.2 1145	0.81	HexA5HexNAc4S1 -B	-1.51	885.210 1	3
B10	5	458.0 604	100.00	HexA5HexNAc5S5 -B	1.21	458.061 0	0
B10	4	552.8 3839	4.50	HexA5HexNAc5S4 -B	0.84	552.838 9	1
B10	3	710.8 0213	4.38	HexA5HexNAc5S3 -B	-0.23	710.802 0	2
B10	4	532.8 4852	0.66	HexA5HexNAc5S3 -B	2.13	532.849 7	2
B12	5	533.8 829	1.05	HexA6HexNAc6S5 -B	0.70	533.883 3	0
B12	4	647.6 1613	27.30	HexA6HexNAc6S4 -B	0.91	647.616 7	1
B12	3	837.1 7294	40.88	HexA6HexNAc6S3 -B	-0.58	837.172 5	2
B12	4	627.6 2676	1.33	HexA6HexNAc6S3 -B	1.21	627.627 5	2
B13	5	569.0 8937	0.77	HexA7HexNAc6S5 -B	0.57	569.089 7	0
B13	4	691.6 2454	0.72	HexA7HexNAc6S4 -B	0.29	691.624 7	1
B13	4	671.6 357	0.75	HexA7HexNAc6S3 -B	-0.23	671.635 5	2
B14	6	521.2 4614	0.30	HexA7HexNAc7S6 -B	0.17	521.246 2	0
B15	6	550.5 8489	0.39	HexA8HexNAc7S6 -B	0.04	550.584 9	0
Y03	3	463.7 6231	2.32	Y-Hex2Pen1-2H	1.27	463.762 9	0
Y04	1	651.1 984	15.30	Y- Hex2HexA1Pen1- 2H	1.15	651.199 1	0
Y05	1	934.2 3641	3.38	Y- Hex2HexA1HexNA c1Pen1S1-2H	-1.51	934.235 0	0
Y05	1	854.2 7917	4.42	Y- Hex2HexA1HexNA c1Pen1-2H	-1.14	854.278 2	1

Y06	2	554.62922	1.72	Y-Hex2HexA2HexNAc1Pen1S1-2H	1.23	554.6299	0
Y06	1	1030.3134	1.07	Y-Hex2HexA2HexNAc1Pen1-2H	-3.02	1030.3103	1
Y07	2	696.14784	4.27	Y-Hex2HexA2HexNAc2Pen1S2-2H	0.23	696.1480	0
Y07	2	656.16922	3.74	Y-Hex2HexA2HexNAc2Pen1S1-2H	0.58	656.1696	1
Y08	2	784.16393	0.50	Y-Hex2HexA3HexNAc2Pen1S2-2H	0.15	784.1640	0
Y08	3	522.43972	0.93	Y-Hex2HexA3HexNAc2Pen1S2-2H	1.05	522.4403	0
Y08	2	744.18608	0.62	Y-Hex2HexA3HexNAc2Pen1S1-2H	-0.58	744.1856	1
Y09	3	616.78484	2.46	Y-Hex2HexA3HexNAc3Pen1S3-2H	1.34	616.7857	0
Y09	4	462.337	0.72	Y-Hex2HexA3HexNAc3Pen1S3-2H	0.92	462.3374	0
Y09	3	590.13285	2.44	Y-Hex2HexA3HexNAc3Pen1S2-2H	0.93	590.1334	1
Y09	3	563.48015	0.28	Y-Hex2HexA3HexNAc3Pen1S1-2H	1.75	563.4811	2
Y10	3	675.46219	0.31	Y-Hex2HexA4HexNAc3Pen1S3-2H	1.25	675.4630	0
Y10	3	648.81123	0.35	Y-Hex2HexA4HexNAc3Pen1S2-2H	-0.71	648.8108	1
Y10	4	466.36736	0.19	Y-Hex2HexA4HexNAc3Pen1S1-2H	-0.66	466.3670	2
Y11	3	743.15485	0.27	Y-Hex2HexA4HexNAc4Pen1S3-2H	1.77	743.1562	0
Y11	4	557.11459	2.32	Y-Hex2HexA4HexNAc4Pen1S3-2H	1.27	557.1153	0
Y12	4	601.12306	0.77	Y-Hex2HexA5HexNAc4Pen1S3-2H	0.44	601.1233	0
Y13	4	671.88186	1.26	Y-Hex2HexA5HexNAc5Pen1S4-2H	0.77	671.8824	0

Y13	5	537.30368	1.39	Y-Hex2HexA5HexNAc5Pen1S4-2H	1.41	537.3044	0
Y13	4	651.89226	0.62	Y-Hex2HexA5HexNAc5Pen1S3-2H	1.40	651.8932	1
Y14	5	572.51064	1.66	Y-Hex2HexA6HexNAc5Pen1S4-2H	0.38	572.5109	0
Y15	5	629.11846	2.44	Y-Hex2HexA6HexNAc6Pen1S5-2H	-0.57	629.1181	0
C02	1	476.0709	3.37	HexA1HexNAc1S1-C	1.36	476.0715	0
C03	1	652.103	0.91	HexA2HexNAc1S1-C	0.99	652.1036	0
C04	2	467.06564	0.64	HexA2HexNAc2S2-C	1.34	467.0663	0
C04	1	855.18391	3.05	HexA2HexNAc1S1-C	-1.03	855.1830	1
C04	1	775.22638	1.42	HexA2HexNAc2-C	-0.19	775.2262	2
C06	3	464.06415	0.41	HexA3HexNAc3S3-C	0.76	464.0645	0
C06	2	656.62148	1.82	HexA3HexNAc3S2-C	0.80	656.6220	1
C06	2	616.64289	0.86	HexA3HexNAc3S1-C	1.16	616.6436	2
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	8.90	HexA2GalNAc3S2	1.25	559.6007	
	2	370.0442	4.80	HexA1GalNAc2S2	2.16	370.0450	
	1	300.0389	1.90	GalNAc1S1-O	2.00	300.0395	
	1	282.0298	0.18	GalNAc1S1	-3.19	282.0289	
	1	661.1399	13.30	HexA1GalNAc2S1	0.76	661.1404	

Supplementary Table 4.17. Assignment of fragment ions resulted from CID of parent ion m/z 586.9750 (z=8) corresponding to composition dp22-8S. CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



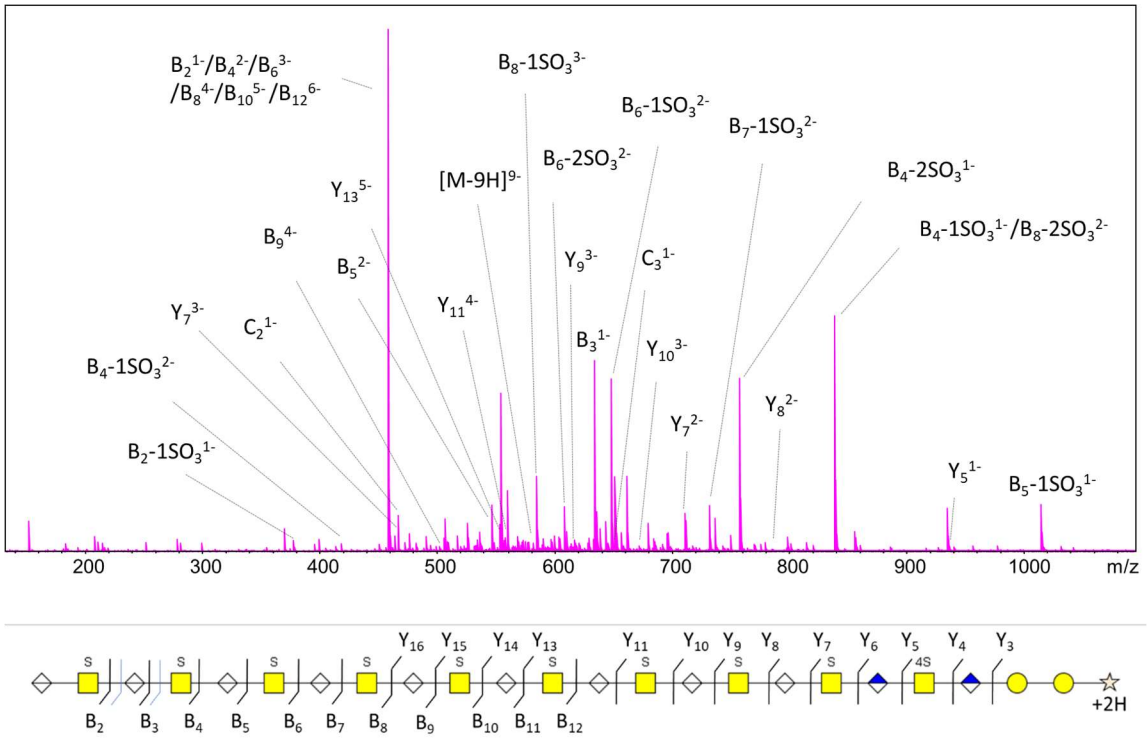
Supplementary Figure 4.26. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 654.3536 (z=8) and its fragmentation pattern providing sequence for composition dp24-10S (f39). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.18.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0606	100.00	HexA1HexNAc1S1-B	0.94	458.0610	0
B03	1	634.0931	41.08	HexA2HexNAc1S1-B	-0.06	634.0931	0
B03	1	554.1357	20.54	HexA2HexNAc1-B	1.07	554.1363	1
B04	2	458.0606	100.00	HexA2HexNAc2S2-B	0.94	458.0610	0
B04	1	837.1735	29.52	HexA2HexNAc2S1-B	-1.27	837.1725	1
B04	1	757.2162	10.94	HexA2HexNAc2-B	-0.68	757.2157	2
B05	2	546.0769	14.74	HexA3HexNAc2S2-B	0.33	546.0770	0
B05	1	1013.2071	4.02	HexA3HexNAc2S1-B	-2.53	1013.2046	1
B05	2	506.0976	2.20	HexA3HexNAc2S1-B	2.00	506.0986	1
B06	3	458.0606	100.00	HexA3HexNAc3S3-B	0.94	458.0610	0
B06	2	647.6167	37.38	HexA3HexNAc3S2-B	0.11	647.6167	1
B06	2	607.6379	4.96	HexA3HexNAc3S1-B	0.65	607.6383	2
B07	3	516.7380	7.34	HexA4HexNAc3S3-B	0.60	516.7383	0
B07	2	735.6336	9.10	HexA4HexNAc3S2-B	-1.06	735.6328	1
B08	4	458.0606	100.00	HexA4HexNAc4S4-B	0.94	458.0610	0
B08	3	584.4311	14.93	HexA4HexNAc4S3-B	0.64	584.4315	1
B08	2	837.1735	29.52	HexA4HexNAc4S2-B	-1.27	837.1725	2
B08	2	797.1964	2.87	HexA4HexNAc4S1-B	-2.97	797.1941	3
B09	3	643.1086	7.44	HexA5HexNAc4S3-B	0.41	643.1088	1
B10	5	458.0606	100.00	HexA5HexNAc5S5-B	0.94	458.0610	0
B10	4	552.8384	4.88	HexA5HexNAc5S4-B	0.75	552.8389	1
B10	3	710.8022	10.11	HexA5HexNAc5S3-B	-0.28	710.8020	2
B12	4	647.6167	37.38	HexA6HexNAc6S4-B	0.11	647.6167	2
B12	3	837.1735	29.52	HexA6HexNAc6S3-B	-1.27	837.1725	3
B14	5	609.7045	2.28	HexA7HexNAc7S5-B	1.70	609.7056	2

B16	6	584.4311	14.93	HexA8HexNAc8S6-B	0.64	584.4315	2
B18	6	647.6167	37.38	HexA9HexNAc9S6-B	0.11	647.6167	3
Y04	1	651.1987	30.40	Y-Hex2HexA1Pen1-2H	0.25	651.1989	0
Y05	1	934.2377	5.69	Y-Hex2HexA1HexNAc1Pen1S1-2H	-2.89	934.2350	0
Y05	2	466.6133	3.61	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.16	466.6138	0
Y06	2	554.6294	6.94	Y-Hex2HexA2HexNAc1Pen1S1-2H	0.87	554.6299	0
Y07	2	696.1485	12.62	Y-Hex2HexA2HexNAc2Pen1S2-2H	-0.75	696.1480	0
Y07	2	656.1695	9.23	Y-Hex2HexA2HexNAc2Pen1S1-2H	0.09	656.1696	1
Y08	2	784.1653	2.14	Y-Hex2HexA3HexNAc2Pen1S2-2H	-1.58	784.1640	0
Y08	3	522.4405	1.52	Y-Hex2HexA3HexNAc2Pen1S2-2H	-0.41	522.4403	0
Y10	3	675.4634	4.17	Y-Hex2HexA4HexNAc3Pen1S3-2H	-0.48	675.4630	0
Y10	3	648.8110	2.05	Y-Hex2HexA4HexNAc3Pen1S2-2H	-0.33	648.8108	1
Y11	4	577.1042	4.08	Y-Hex2HexA4HexNAc4Pen1S4-2H	0.54	577.1045	0
Y11	3	743.1569	3.13	Y-Hex2HexA4HexNAc4Pen1S3-2H	-0.96	743.1562	1
Y11	4	557.1141	1.88	Y-Hex2HexA4HexNAc4Pen1S3-2H	2.21	557.1153	1
Y12	4	621.1123	3.39	Y-Hex2HexA5HexNAc4Pen1S4-2H	0.44	621.1125	0
Y15	5	629.1172	2.66	Y-Hex2HexA6HexNAc6Pen1S6-2H	1.46	629.1181	1
C02	1	476.0711	4.40	HexA1HexNAc1S1-C	1.00	476.0715	0
C03	1	652.1035	2.66	HexA2HexNAc1S1-C	0.26	652.1036	0

C04	2	467.0655	1.41	HexA2HexNAc2S2-C	1.74	467.0663	0
C05	2	555.0820	1.55	HexA3HexNAc2S2-C	0.53	555.0823	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6001	12.90	HexA2GalNAc3S2	1.07	559.6007	
	2	370.0444	6.68	HexA1GalNAc2S2	1.62	370.0450	
	1	300.0390	12.70	GalNAc1S1-O	1.67	300.0395	
	1	282.0285	1.20	GalNAc1S1	1.42	282.0289	
	1	661.1402	17.40	HexA1GalNAc2S1	0.30	661.1404	

Supplementary Table 4.18. Assignment of fragment ions resulted from CID of parent ion m/z 654.3536 (z=8) corresponding to composition dp24-10S (f39). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



Supplementary Figure 4.28. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 581.5353 (z=9) and its fragmentation pattern providing sequence for composition dp24-10S (f39). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.19.

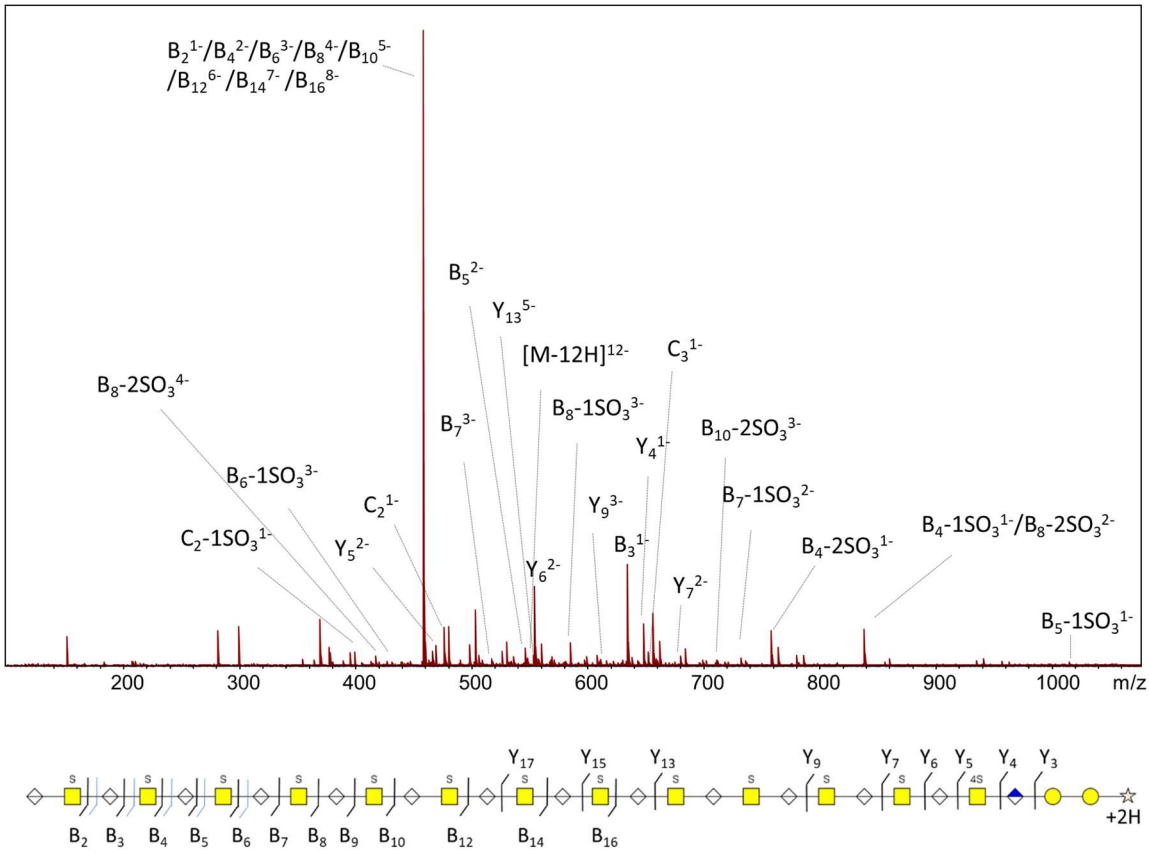
CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0605	100.00	HexA1HexNAc1S1-B	1.10	458.0610	0
B02	1	378.1035	2.16	HexA1HexNAc1-B	1.83	378.1042	1
B03	1	634.0930	36.59	HexA2HexNAc1S1-B	0.21	634.0931	0
B04	2	458.0605	100.00	HexA2HexNAc2S2-B	1.10	458.0610	0
B04	1	837.1731	45.28	HexA2HexNAc2S1-B	-0.74	837.1725	1
B04	2	418.0820	1.45	HexA2HexNAc2S1-B	1.46	418.0826	1
B04	1	757.2161	33.11	HexA2HexNAc2-B	-0.60	757.2157	2
B04	2	378.1035	2.16	HexA2HexNAc2-B	1.83	378.1042	2
B05	2	546.0767	8.94	HexA3HexNAc2S2-B	0.66	546.0770	0
B05	1	1013.2073	9.03	HexA3HexNAc2S1-B	-2.74	########	1
B05	2	506.0980	2.60	HexA3HexNAc2S1-B	1.21	506.0986	1
B05	1	933.2498	8.27	HexA3HexNAc2-B	-2.17	933.2478	2
B06	3	458.0605	100.00	HexA3HexNAc3S3-B	1.10	458.0610	0
B06	2	647.6163	33.09	HexA3HexNAc3S2-B	0.71	647.6167	1
B06	3	431.4078	0.32	HexA3HexNAc3S2-B	2.10	431.4087	1
B06	2	607.6378	8.53	HexA3HexNAc3S1-B	0.83	607.6383	2
B06	3	378.1035	2.16	HexA3HexNAc3-B	1.83	378.1042	3
B07	3	516.7379	2.98	HexA4HexNAc3S3-B	0.93	516.7383	0
B07	2	735.6328	6.40	HexA4HexNAc3S2-B	-0.09	735.6328	1
B07	3	490.0858	0.57	HexA4HexNAc3S2-B	0.64	490.0861	1
B07	2	695.6547	3.47	HexA4HexNAc3S1-B	-0.51	695.6544	2
B08	4	458.0605	100.00	HexA4HexNAc4S4-B	1.10	458.0610	0
B08	3	584.4309	14.41	HexA4HexNAc4S3-B	1.00	584.4315	1
B08	2	837.1731	45.28	HexA4HexNAc4S2-B	-0.74	837.1725	2
B08	3	557.7787	2.68	HexA4HexNAc4S2-B	0.95	557.7792	2

B08	4	418.0820	1.45	HexA4HexNAc4S2-B	1.46	418.0826	2
B08	2	797.1947	2.79	HexA4HexNAc4S1-B	-0.84	797.1941	3
B09	4	502.0684	0.92	HexA5HexNAc4S4-B	1.23	502.0690	0
B09	3	643.1086	5.76	HexA5HexNAc4S3-B	0.36	643.1088	1
B09	2	925.1913	0.70	HexA5HexNAc4S2-B	-3.05	925.1885	2
B09	3	616.4562	1.79	HexA5HexNAc4S2-B	0.56	616.4566	2
B09	2	885.2113	0.83	HexA5HexNAc4S1-B	-1.38	885.2101	3
B10	5	458.0605	100.00	HexA5HexNAc5S5-B	1.10	458.0610	0
B10	4	552.8385	5.24	HexA5HexNAc5S4-B	0.73	552.8389	1
B10	3	710.8022	7.43	HexA5HexNAc5S3-B	-0.38	710.8020	2
B10	4	532.8491	1.15	HexA5HexNAc5S3-B	1.07	532.8497	2
B11	5	493.2672	0.20	HexA6HexNAc5S5-B	0.47	493.2674	0
B11	4	596.8464	2.22	HexA6HexNAc5S4-B	0.83	596.8469	1
B11	3	769.4794	1.36	HexA6HexNAc5S3-B	-0.06	769.4793	2
B11	4	576.8572	0.58	HexA6HexNAc5S3-B	0.86	576.8577	2
B12	6	458.0605	100.00	HexA6HexNAc6S6-B	1.10	458.0610	0
B12	4	647.6163	33.09	HexA6HexNAc6S4-B	0.71	647.6167	2
Y03	1	475.1663	0.57	Y-Hex2Pen1-2H	1.12	475.1668	0
Y04	1	651.1985	14.26	Y-Hex2HexA1Pen1-2H	0.54	651.1989	0
Y05	1	934.2358	2.72	Y-Hex2HexA1HexNAc1Pen1S1-2H	-0.90	934.2350	0
Y05	2	466.6133	6.86	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.26	466.6138	0
Y05	1	854.2796	3.85	Y-Hex2HexA1HexNAc1Pen1-2H	-1.59	854.2782	1
Y06	2	554.6294	1.53	Y-Hex2HexA2HexNAc1Pen1S1-2H	0.87	554.6299	0
Y06	1	1030.3132	0.97	Y-Hex2HexA2HexNAc1Pen1-2H	-2.80	########	1

Y07	2	696.1480	3.66	Y-Hex2HexA2HexNAc2Pen1S2-2H	-0.03	696.1480	0
Y07	3	463.7623	2.99	Y-Hex2HexA2HexNAc2Pen1S2-2H	1.25	463.7629	0
Y07	2	656.1692	3.58	Y-Hex2HexA2HexNAc2Pen1S1-2H	0.56	656.1696	1
Y08	2	784.1651	0.29	Y-Hex2HexA3HexNAc2Pen1S2-2H	-1.34	784.1640	0
Y08	3	522.4397	1.01	Y-Hex2HexA3HexNAc2Pen1S2-2H	1.01	522.4403	0
Y08	2	744.1857	0.51	Y-Hex2HexA3HexNAc2Pen1S1-2H	-0.07	744.1856	1
Y09	3	616.7849	2.20	Y-Hex2HexA3HexNAc3Pen1S3-2H	1.31	616.7857	0
Y09	4	462.3370	0.71	Y-Hex2HexA3HexNAc3Pen1S3-2H	1.01	462.3374	0
Y09	3	590.1329	2.44	Y-Hex2HexA3HexNAc3Pen1S2-2H	0.80	590.1334	1
Y10	3	675.4635	0.34	Y-Hex2HexA4HexNAc3Pen1S3-2H	-0.63	675.4630	0
Y10	4	506.3449	0.41	Y-Hex2HexA4HexNAc3Pen1S3-2H	1.18	506.3454	0
Y11	4	577.1037	1.70	Y-Hex2HexA4HexNAc4Pen1S4-2H	1.33	577.1045	0
Y11	3	716.5035	0.47	Y-Hex2HexA4HexNAc4Pen1S2-2H	0.61	716.5039	2
Y13	5	553.2952	0.59	Y-Hex2HexA5HexNAc5Pen1S5-2H	1.03	553.2958	0
Y14	5	588.5018	0.57	Y-Hex2HexA6HexNAc5Pen1S5-2H	0.75	588.5022	0
Y15	6	537.4221	0.38	Y-Hex2HexA6HexNAc6Pen1S6-2H	2.26	537.4233	0
Y16	6	566.7614	0.61	Y-Hex2HexA7HexNAc6Pen1S6-2H	1.14	566.7620	0
C02	1	476.0711	3.44	HexA1HexNAc1S1-C	1.04	476.0715	0

C03	1	652.1034	1.08	HexA2HexNAc1S1-C	0.32	652.1036	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.6000	11.80	HexA2GalNAc3S2	1.25	559.6007	
	2	370.0443	4.41	HexA1GalNAc2S2	1.89	370.0450	
	1	300.0389	1.55	GalNAc1S1-O	2.00	300.0395	
	1	282.0284	1.60	GalNAc1S1	1.77	282.0289	
	1	661.1400	14.60	HexA1GalNAc2S1	0.61	661.1404	

Supplementary Table 4.19. Assignment of fragment ions resulted from CID of parent ion m/z 581.5353 (z=9) corresponding to composition dp24-10S (f39). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



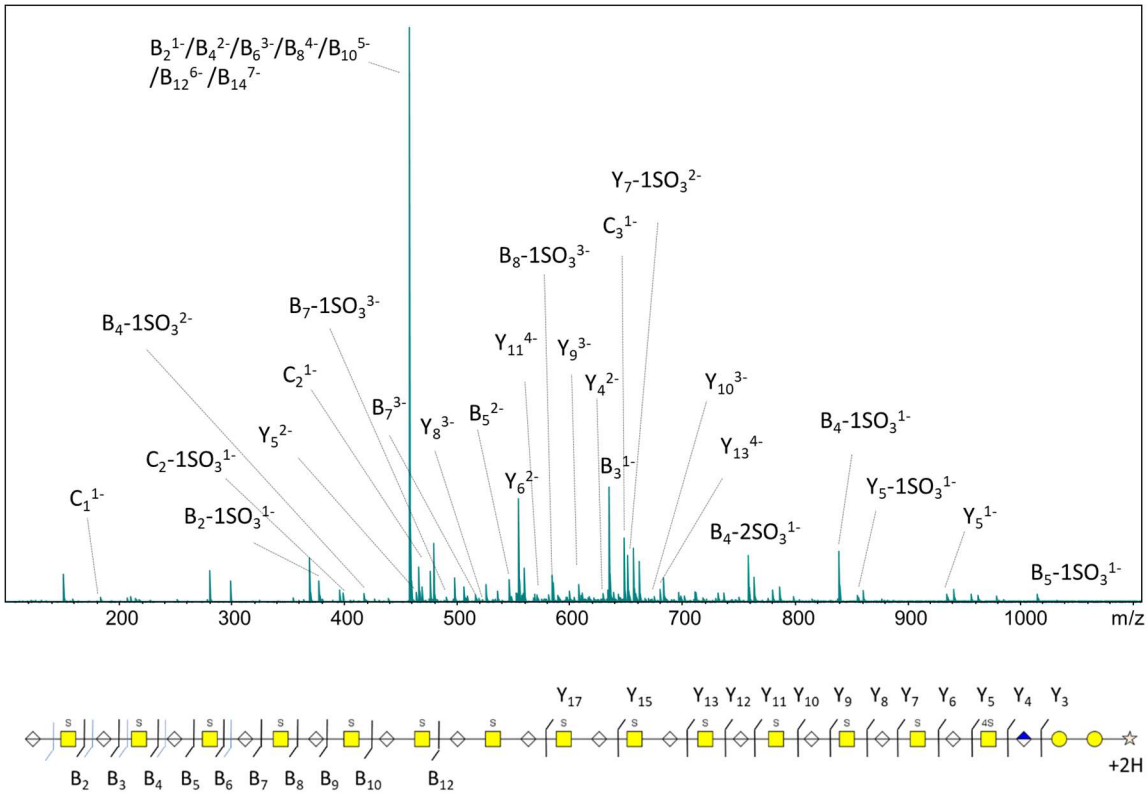
Supplementary Figure 4.29. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 550.6652 (z=12) and its fragmentation pattern providing sequence for composition dp30-13S (f51). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.20.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0598	100.00	HexA1HexNAc1S1-B	2.69	458.0610	0
B03	1	634.0917	15.73	HexA2HexNAc1S1-B	2.24	634.0931	0
B03	2	316.5420	0.29	HexA2HexNAc1S1-B	2.69	316.5429	0
B03	1	554.1389	2.69	HexA2HexNAc1-B	####	554.1363	1
B04	2	458.0598	100.00	HexA2HexNAc2S2-B	2.69	458.0610	0
B04	1	837.1719	5.71	HexA2HexNAc2S1-B	0.66	837.1725	1
B04	2	418.0814	1.60	HexA2HexNAc2S1-B	2.95	418.0826	1
B04	1	757.2149	5.53	HexA2HexNAc2-B	0.96	757.2157	2
B05	2	546.0757	2.81	HexA3HexNAc2S2-B	2.49	546.0770	0
B05	1	1013.2053	0.66	HexA3HexNAc2S1-B	####	1013.2046	1
B05	2	506.0974	1.13	HexA3HexNAc2S1-B	2.43	506.0986	1
B06	3	458.0598	100.00	HexA3HexNAc3S3-B	2.69	458.0610	0
B06	3	431.4077	0.63	HexA3HexNAc3S2-B	2.31	431.4087	1
B06	2	607.6374	1.72	HexA3HexNAc3S1-B	1.49	607.6383	2
B07	3	516.7369	1.14	HexA4HexNAc3S3-B	2.76	516.7383	0
B07	2	735.6311	0.85	HexA4HexNAc3S2-B	2.22	735.6328	1
B07	3	490.0846	0.42	HexA4HexNAc3S2-B	2.96	490.0861	1
B07	2	695.6529	0.48	HexA4HexNAc3S1-B	2.10	695.6544	2
B08	4	458.0598	100.00	HexA4HexNAc4S4-B	2.69	458.0610	0
B08	3	584.4302	3.67	HexA4HexNAc4S3-B	2.13	584.4315	1
B08	2	837.1719	5.71	HexA4HexNAc4S2-B	0.66	837.1725	2
B08	3	557.7781	0.81	HexA4HexNAc4S2-B	2.02	557.7792	2
B08	4	418.0814	1.60	HexA4HexNAc4S2-B	2.95	418.0826	2
B09	4	502.0676	0.66	HexA5HexNAc4S4-B	2.90	502.0690	0
B09	3	643.1073	0.85	HexA5HexNAc4S3-B	2.35	643.1088	1

B09	3	616.4541	0.34	HexA5HexNAc4S2-B	3.97	616.4566	2
B10	5	458.0598	100.00	HexA5HexNAc5S5-B	2.69	458.0610	0
B10	4	552.8378	1.59	HexA5HexNAc5S4-B	1.87	552.8389	1
B10	3	710.8007	0.90	HexA5HexNAc5S3-B	1.77	710.8020	2
B10	4	532.8484	0.61	HexA5HexNAc5S3-B	2.39	532.8497	2
B12	6	458.0598	100.00	HexA6HexNAc6S6-B	2.69	458.0610	0
B12	5	533.8820	0.74	HexA6HexNAc6S5-B	2.32	533.8833	1
B12	6	431.4077	0.63	HexA6HexNAc6S4-B	2.31	431.4087	2
B14	7	458.0598	100.00	HexA7HexNAc7S7-B	2.69	458.0610	0
B14	5	609.7041	0.50	HexA7HexNAc7S5-B	2.42	609.7056	2
B16	8	458.0598	100.00	HexA8HexNAc8S8-B	2.69	458.0610	0
B16	6	584.4302	3.67	HexA8HexNAc8S6-B	2.13	584.4315	2
B20	8	552.8378	1.59	HexA10HexNAc10S8-B	1.87	552.8389	2
B24	10	533.8820	0.74	HexA12HexNAc12S10-B	2.32	533.8833	2
Y03	1	475.1655	0.27	Y-Hex2Pen1-2H	3.27	475.1671	0
Y04	1	651.1976	2.14	Y-Hex2HexA1Pen1-2H	2.38	651.1991	0
Y05	2	466.6129	2.26	Y-Hex2HexA1HexNAc1Pen1S1-2H	2.10	466.6138	0
Y06	2	554.6288	0.33	Y-Hex2HexA2HexNAc1Pen1S1-2H	2.02	554.6299	0
Y07	2	696.1465	0.41	Y-Hex2HexA2HexNAc2Pen1S2-2H	2.21	696.1480	0
Y07	3	463.7618	0.92	Y-Hex2HexA2HexNAc2Pen1S2-2H	2.37	463.7629	0
Y09	3	616.7840	0.28	Y-Hex2HexA3HexNAc3Pen1S3-2H	2.72	616.7857	0
Y09	4	462.3363	0.39	Y-Hex2HexA3HexNAc3Pen1S3-2H	2.41	462.3374	0
Y13	5	553.2944	0.31	Y-Hex2HexA5HexNAc5Pen1S5-2H	2.53	553.2958	0

Y13	5	537.3031	0.47	Y-Hex2HexA5HexNAc5Pen1S4-2H	2.46	537.3044	1
Y15	5	645.1114	0.28	Y-Hex2HexA6HexNAc6Pen1S6-2H	####	645.1095	0
Y15	6	537.4222	0.30	Y-Hex2HexA6HexNAc6Pen1S6-2H	2.05	537.4233	0
Y17	7	526.0873	1.29	Y-Hex2HexA7HexNAc7Pen1S7-2H	####	526.0859	0
C02	1	476.0704	6.01	HexA1HexNAc1S1-C	2.33	476.0715	0
C02	1	396.1136	2.09	HexA1HexNAc1-C	2.85	396.1147	1
C03	1	652.1023	0.69	HexA2HexNAc1S1-C	2.13	652.1036	0
C04	2	467.0648	0.99	HexA2HexNAc2S2-C	3.07	467.0663	0
C05	2	555.0808	0.29	HexA3HexNAc2S2-C	2.67	555.0823	0
C06	3	464.0628	0.25	HexA3HexNAc3S3-C	3.58	464.0645	0
C06	2	656.6213	0.62	HexA3HexNAc3S2-C	1.15	656.6220	1
C06	2	616.6420	0.35	HexA3HexNAc3S1-C	2.62	616.6436	2
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.5994	3.45	HexA2GalNAc3S2	2.32	559.6007	
	2	370.0439	7.32	HexA1GalNAc2S2	2.97	370.0450	
	1	300.0387	6.19	GalNAc1S1-O	2.67	300.0395	
	1	282.0283	5.56	GalNAc1S1	2.13	282.0289	
	1	661.1394	3.86	HexA1GalNAc2S1	1.51	661.1404	

Supplementary Table 4.20. Assignment of fragment ions resulted from CID of parent ion m/z 550.6652 (z=12) corresponding to composition dp30-13S (f51). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



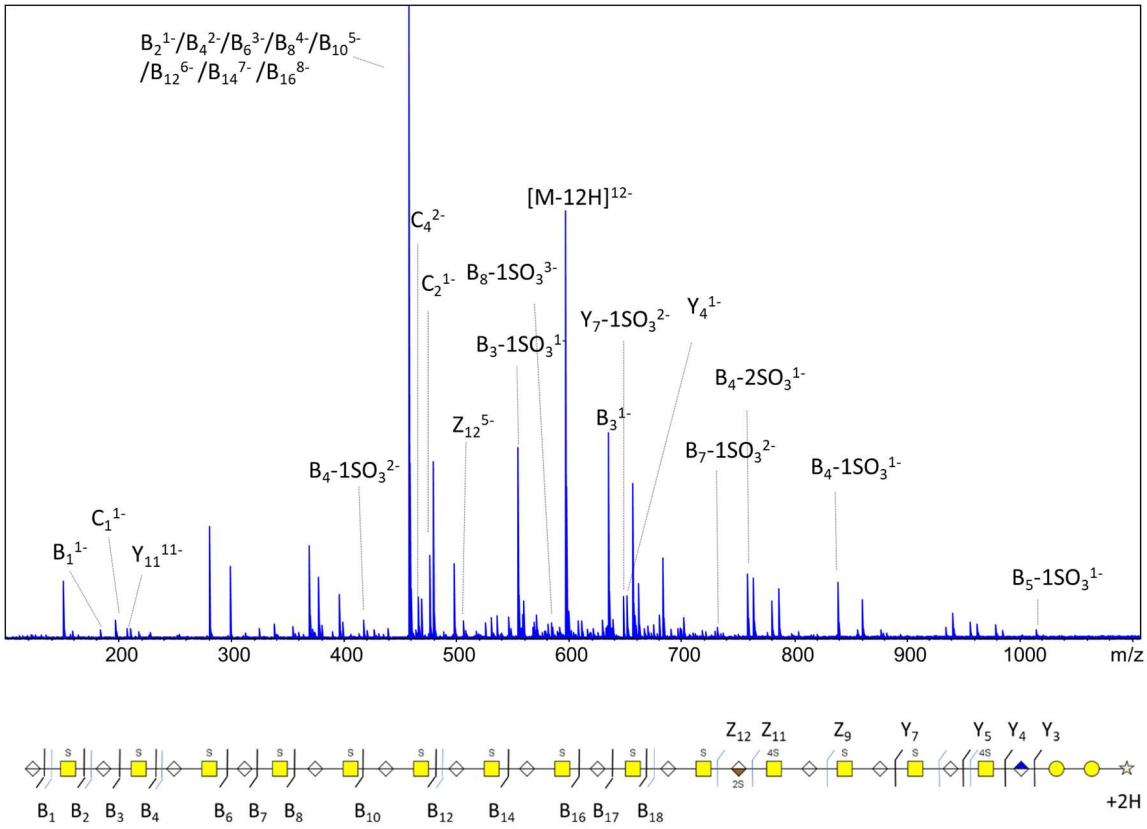
Supplementary Figure 4.30. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 642.5506 (z=11) and its fragmentation pattern providing sequence for composition dp32-14S (f51). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.21.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0597	100.00	HexA1HexNAc1S1-B	2.75	458.0610	0
B02	1	378.1058	3.59	HexA1HexNAc1-B	-4.40	378.1042	1
B03	1	634.0916	19.69	HexA2HexNAc1-B	2.30	634.0931	0
B04	2	458.0597	100.00	HexA2HexNAc2S2-B	2.75	458.0610	0
B04	1	837.1717	8.83	HexA2HexNAc2S1-B	0.89	837.1724	1
B04	2	418.0825	1.46	HexA2HexNAc2S1-B	0.22	418.0826	1
B04	1	757.2178	7.95	HexA2HexNAc2-B	-2.87	757.2157	2
B04	2	378.1058	3.59	HexA2HexNAc2-B	-4.40	378.1042	2
B05	2	546.0793	3.90	HexA3HexNAc2S2-B	-4.22	546.0770	0
B05	1	1013.2056	1.32	HexA3HexNAc2S1-B	-1.05	1013.2046	1
B06	3	458.0597	100.00	HexA3HexNAc3S3-B	2.75	458.0610	0
B06	3	378.1058	3.59	HexA3HexNAc3-B	-4.40	378.1042	3
B07	3	516.7368	1.23	HexA4HexNAc3S3-B	2.94	516.7383	0
B07	3	490.0850	0.24	HexA4HexNAc3S2-B	2.15	490.0861	1
B07	2	695.6530	0.84	HexA4HexNAc3S1-B	1.93	695.6544	2
B08	4	458.0597	100.00	HexA4HexNAc4S4-B	2.75	458.0610	0
B08	3	584.4314	4.59	HexA4HexNAc4S3-B	0.12	584.4315	1
B08	3	557.7780	1.16	HexA4HexNAc4S2-B	2.17	557.7792	2
B08	4	418.0825	1.46	HexA4HexNAc4S2-B	0.22	418.0826	2
B08	2	797.1938	0.90	HexA4HexNAc4S1-B	0.36	797.1941	3
B10	5	458.0597	100.00	HexA5HexNAc5S5-B	2.75	458.0610	0
B10	4	552.8379	1.58	HexA5HexNAc5S4-B	1.78	552.8389	1
B10	3	710.8007	1.76	HexA5HexNAc5S3-B	1.77	710.8020	2
B12	6	458.0597	100.00	HexA6HexNAc6S6-B	2.75	458.0610	0
B12	5	533.8818	0.45	HexA6HexNAc6S5-B	2.81	533.8833	1

B14	5	609.7041	1.01	HexA7HexNAc6S5-B	2.35	609.7056	2
Y03	1	475.1654	0.55	Y-Hex2Pen1-2H	3.52	475.1671	0
Y04	1	651.1976	7.98	Y-Hex2HexA1Pen1-2H	2.49	651.1992	0
Y05	1	934.2333	0.78	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.84	934.2350	0
Y05	2	466.6131	6.01	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.54	466.6138	0
Y05	1	854.2773	1.17	Y-Hex2HexA1HexNAc1Pen1-2H	1.09	854.2782	1
Y06	2	554.6289	0.88	Y-Hex2HexA2HexNAc1Pen1S1-2H	1.80	554.6299	0
Y07	3	463.7618	1.65	Y-Hex2HexA2HexNAc2Pen1S2-2H	2.48	463.7629	0
Y07	2	656.1673	2.68	Y-Hex2HexA2HexNAc2Pen1S1-2H	3.44	656.1696	1
Y08	3	522.4393	0.39	Y-Hex2HexA3HexNAc2Pen1S2-2H	1.81	522.4403	0
Y09	3	616.7840	0.90	Y-Hex2HexA3HexNAc3Pen1S3-2H	2.75	616.7857	0
Y09	4	462.3362	0.30	Y-Hex2HexA3HexNAc3Pen1S3-2H	2.61	462.3374	0
Y09	3	590.1319	0.85	Y-Hex2HexA3HexNAc3Pen1S2-2H	2.58	590.1334	1
Y10	3	675.4623	0.28	Y-Hex2HexA4HexNAc3Pen1S2-2H	1.16	675.4630	0
Y11	4	577.1033	0.43	Y-Hex2HexA4HexNAc4Pen1S4-2H	2.15	577.1045	0
Y11	3	743.1534	0.36	Y-Hex2HexA4HexNAc4Pen1S3-2H	3.74	743.1562	1
Y11	4	557.1163	0.81	Y-Hex2HexA4HexNAc4Pen1S3-2H	-1.71	557.1153	1
Y12	4	621.1112	0.33	Y-Hex2HexA5HexNAc4Pen1S4-2H	2.13	621.1125	0

Y13	4	691.8735	0.34	Y-Hex2HexA5HexNAc5Pen1S5-2H	-2.74	691.8716	0
Y13	5	553.2940	0.25	Y-Hex2HexA5HexNAc5Pen1S5-2H	3.27	553.2958	0
Y13	4	671.8800	0.40	Y-Hex2HexA5HexNAc5Pen1S4-2H	3.55	671.8824	1
Y15	5	645.1118	0.41	Y-Hex2HexA6HexNAc6Pen1S6-2H	-3.61	645.1095	0
Y15	5	629.1177	0.74	Y-Hex2HexA6HexNAc6Pen1S5-2H	0.72	629.1181	1
Y17	6	613.9339	0.48	Y-Hex2HexA7HexNAc7Pen1S7-2H	1.27	613.9347	0
C01	1	193.0351	0.14	HexA1-C	1.37	193.0354	0
C02	1	476.0703	5.30	HexA1HexNAc1S1-C	2.52	476.0715	0
C02	1	396.1136	2.12	HexA1HexNAc1-C	2.86	396.1147	1
C03	1	652.1045	1.44	HexA2HexNAc1S1-C	-1.24	652.1036	0
C04	2	467.0648	0.48	HexA2HexNAc2S2-C	3.24	467.0663	0
C06	3	464.0630	0.21	HexA3HexNAc3S3-C	3.28	464.0645	0
C06	2	656.6207	0.53	HexA3HexNAc3S2-C	2.05	656.6220	1
C12	6	461.0631	0.52	HexA6HexNAc6S6-C	-0.87	461.0627	0
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.5994	5.80	HexA2GalNAc3S2	2.32	559.6007	
	2	370.0439	7.57	HexA1GalNAc2S2	2.97	370.0450	
	1	300.0399	0.82	GalNAc1S1-O	-1.33	300.0395	
	1	282.0283	5.40	GalNAc1S1	2.13	282.0289	
	1	661.1394	6.90	HexA1GalNAc2S1	1.51	661.1404	

Supplementary Table 4.21. Assignment of fragment ions resulted from CID of parent ion m/z 642.5506 (z=11) corresponding to composition dp32-14S (f51). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



Supplementary Figure 4.31. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 595.5842 (z=12) and its fragmentation pattern providing sequence for composition dp32-15S (f51). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.22.

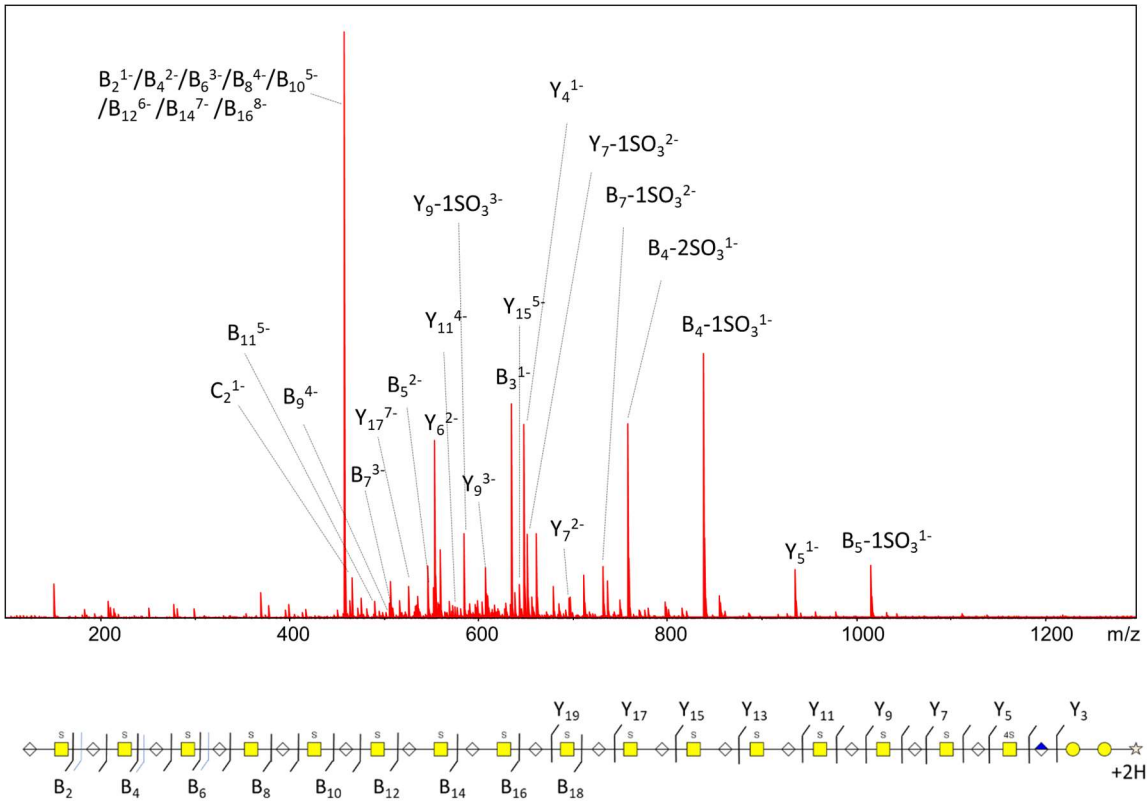
CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B01	1	175.0246	0.20	HexA1-B	1.20	175.0248	0
B02	1	458.0597	100.00	HexA1HexNAc1S1-B	2.76	458.0610	0
B03	1	634.0916	16.16	HexA2HexNAc1S1-B	2.35	634.0931	0
B03	1	554.1389	3.22	HexA2HexNAc1-B	-4.80	554.1363	1
B04	2	458.0597	100.00	HexA2HexNAc2S2-B	2.76	458.0610	0
B04	1	837.1718	4.51	HexA2HexNAc2S1-B	0.85	837.1725	1
B04	2	418.0813	1.45	HexA2HexNAc2S1-B	2.99	418.0826	1
B04	1	757.2149	5.09	HexA2HexNAc2-B	1.02	757.2157	2
B05	1	1013.2060	0.71	HexA3HexNAc2S1-B	-1.42	########	1
B05	2	506.0976	0.76	HexA3HexNAc2S1-B	2.00	506.0986	1
B06	3	458.0597	100.00	HexA3HexNAc3S3-B	2.76	458.0610	0
B06	2	607.6373	1.43	HexA3HexNAc3S1-B	1.72	607.6383	2
B07	3	516.7369	0.27	HexA4HexNAc3S3-B	2.88	516.7383	0
B07	2	735.6320	0.27	HexA4HexNAc3S2-B	1.09	735.6328	1
B07	3	490.0850	0.24	HexA4HexNAc3S2-B	2.21	490.0861	1
B07	2	695.6534	0.44	HexA4HexNAc3S1-B	1.43	695.6544	2
B08	4	458.0597	100.00	HexA4HexNAc4S4-B	2.76	458.0610	0
B08	3	584.4301	1.26	HexA4HexNAc4S3-B	2.30	584.4315	1
B08	2	837.1718	4.51	HexA4HexNAc4S2-B	0.85	837.1725	2
B08	3	557.7782	0.51	HexA4HexNAc4S2-B	1.76	557.7792	2
B08	4	418.0813	1.45	HexA4HexNAc4S2-B	2.99	418.0826	2
B10	5	458.0597	100.00	HexA5HexNAc5S5-B	2.76	458.0610	0
B12	6	458.0597	100.00	HexA6HexNAc6S6-B	2.76	458.0610	0
B12	3	837.1718	4.51	HexA6HexNAc6S3-B	0.85	837.1725	3
B12	6	418.0813	1.45	HexA6HexNAc6S3-B	2.99	418.0826	3

B14	7	458.0597	100.00	HexA7HexNAc7S7-B	2.76	458.0610	0
B16	8	458.0597	100.00	HexA8HexNAc8S8-B	2.76	458.0610	0
B17	8	480.0647	1.05	HexA9HexNAc8S8-B	0.64	480.0650	0
B18	9	458.0597	100.00	HexA9HexNAc9S9-B	2.76	458.0610	0
Y03	1	475.1655	0.38	Y-Hex2Pen1-2H	3.42	475.1671	0
Y04	1	651.1976	3.35	Y-Hex2HexA1Pen1-2H	2.42	651.1992	0
Y05	2	466.6129	3.26	Y-Hex2HexA1HexNAc1Pen1S1-2H	2.51	466.6141	0
Y05	1	854.2776	0.32	Y-Hex2HexA1HexNAc1Pen1-2H	1.15	854.2786	1
Y07	3	463.7617	0.72	Y-Hex2HexA2HexNAc2Pen1S2-2H	2.63	463.7629	0
Y07	2	656.1692	0.54	Y-Hex2HexA2HexNAc2Pen1S1-2H	0.69	656.1696	1
Y11	11	209.2149	0.54	Y-Hex2HexA4HexNAc4Pen1S4-2H	1.50	209.2152	0
C01	1	193.0351	0.17	HexA1-C	1.32	193.0354	0
C02	1	476.0707	6.50	HexA1HexNAc1S1-C	1.73	476.0715	0
C04	2	467.0649	0.62	HexA2HexNAc2S2-C	2.84	467.0663	0
C04	1	855.1824	0.68	HexA2HexNAc2S1-C	0.79	855.1830	1
C04	2	427.0865	0.72	HexA2HexNAc2S1-C	3.19	427.0879	1
C12	6	461.0633	0.60	HexA6HexNAc6S6-C	-1.13	461.0627	0
C18	9	460.0593	3.91	HexA9HexNAc9S9-C	6.23	460.0622	0
Z05	2	457.6064	0.23	Z-Hex2HexA1HexNAc1Pen1S1-2H	4.84	457.6086	0
Z06	3	457.7568	0.35	Z-Hex2HexA2HexNAc1Pen1S1-2H	3.65	457.7585	0
Z09	4	457.8358	0.36	Z-Hex2HexA3HexNAc3Pen1S3-2H	-2.12	457.8348	0
Z11	5	457.8804	0.44	Z-Hex2HexA4HexNAc4Pen1S4-2H	-0.80	457.8800	0

Z12	5	509.0804	0.24	Z-Hex2HexA5HexNAc4Pen1S5-2H	-4.97	509.0778	0

INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**
	2	559.5995	2.97	HexA2GalNAc3S2	2.14	559.6007
	2	370.0439	7.37	HexA1GalNAc2S2	2.97	370.0450
	1	300.0387	5.70	GalNAc1S1-O	2.67	300.0395
	1	282.0283	8.88	GalNAc1S1	2.13	282.0289
	1	661.1393	4.30	HexA1GalNAc2S1	1.66	661.1404

Supplementary Table 4.22. Assignment of fragment ions resulted from CID of parent ion m/z 595.5842 (z=12) corresponding to composition dp32-15S (f51). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer.



Supplementary Figure 4.32. Annotated spectra from CID-FT-ICR MS/MS in negative ion mode of decorin GAG parent ion m/z 570.2252 (z=14) and its fragmentation pattern providing sequence for composition dp36-16S (f51). Structure diagram appears with relevant B_n and Y_n fragments in black, C_n and Z_n fragments in blue. All fragment ions assigned are provided in Supplementary Table 4.23.

CLEAVAGE	CHARGE	M/Z	RELATIVE INTENSITY	ION	MASS ERROR (PPM)	M/Z (THEOR.)	#SO3 LOSS
B02	1	458.0597	100.00	HexA1HexNAc1S1-B	2.73	458.0610	0
B03	1	634.0924	24.23	HexA2HexNAc1S1-B	1.10	634.0931	0
B04	2	458.0597	100.00	HexA2HexNAc2S2-B	2.73	458.0610	0
B04	1	757.2148	3.80	HexA2HexNAc2-B	1.13	757.2157	2
B05	2	546.0756	6.80	HexA3HexNAc2S2-B	2.71	546.0770	0
B05	1	1013.2073	1.32	HexA3HexNAc2S1-B	-0.99	1013.2063	1
B05	2	506.0973	1.76	HexA3HexNAc2S1-B	2.55	506.0986	1
B06	3	458.0597	100.00	HexA3HexNAc3S3-B	2.73	458.0610	0
B06	2	647.6164	12.94	HexA3HexNAc3S2-B	0.45	647.6167	1
B06	2	607.6376	1.32	HexA3HexNAc3S1-B	1.24	607.6383	2
B07	3	516.7368	4.96	HexA4HexNAc3S3-B	3.07	516.7383	0
B07	2	735.6317	1.91	HexA4HexNAc3S2-B	1.42	735.6328	1
B07	2	695.6532	0.79	HexA4HexNAc3S1-B	1.64	695.6544	2
B08	4	458.0597	100.00	HexA4HexNAc4S4-B	2.73	458.0610	0
B09	4	502.0676	3.26	HexA5HexNAc4S4-B	2.72	502.0690	0
B09	3	643.1072	2.62	HexA5HexNAc4S3-B	2.57	643.1088	1
B09	3	616.4551	0.84	HexA5HexNAc4S2-B	2.38	616.4566	2
B10	5	458.0597	100.00	HexA5HexNAc5S5-B	2.73	458.0610	0
B10	3	710.8003	2.01	HexA5HexNAc5S3-B	2.29	710.8020	2
B11	5	493.2666	0.89	HexA6HexNAc5S5-B	1.61	493.2674	0
B11	4	596.8454	2.54	HexA6HexNAc5S4-B	2.41	596.8469	1
B12	6	458.0597	100.00	HexA6HexNAc6S6-B	2.73	458.0610	0
B12	5	533.8819	2.46	HexA6HexNAc6S5-B	2.56	533.8833	1
B12	4	647.6164	12.94	HexA6HexNAc6S4-B	0.45	647.6167	2
B14	7	458.0597	100.00	HexA7HexNAc7S7-B	2.73	458.0610	0

B14	5	609.7034	1.75	HexA7HexNAc7S5-B	3.58	609.7056	2
B16	8	458.0597	100.00	HexA8HexNAc8S8-B	2.73	458.0610	0
B18	9	458.0597	100.00	HexA9HexNAc9S9-B	2.73	458.0610	0
B18	7	566.3772	1.31	HexA9HexNAc9S7-B	2.36	566.3785	2
B18	6	647.6164	12.94	HexA9HexNAc9S6-B	0.45	647.6167	3
B24	10	533.8819	2.46	HexA12HexNAc12S10-B	2.56	533.8833	2
B28	12	521.2449	1.01	HexA14HexNAc14S12-B	2.51	521.2462	2
Y03	1	475.1650	0.61	Y-Hex2Pen1-2H	4.40	475.1671	0
Y04	1	651.1976	7.66	Y-Hex2HexA1Pen1-2H	0.24	651.1978	0
Y05	1	934.2337	0.84	Y-Hex2HexA1HexNAc1Pen1S1-2H	1.35	934.2350	0
Y05	2	466.6128	7.64	Y-Hex2HexA1HexNAc1Pen1S1-2H	2.34	466.6138	0
Y06	2	554.6285	1.56	Y-Hex2HexA2HexNAc1Pen1S1-2H	2.45	554.6299	0
Y07	2	696.1467	1.92	Y-Hex2HexA2HexNAc2Pen1S2-2H	1.84	696.1480	0
Y07	3	463.7617	3.79	Y-Hex2HexA2HexNAc2Pen1S2-2H	2.57	463.7629	0
Y07	2	656.1686	1.66	Y-Hex2HexA2HexNAc2Pen1S1-2H	1.46	656.1696	1
Y08	3	522.4392	1.29	Y-Hex2HexA3HexNAc2Pen1S2-2H	2.14	522.4403	0
Y09	3	616.7838	2.22	Y-Hex2HexA3HexNAc3Pen1S3-2H	3.06	616.7857	0
Y09	4	462.3360	1.81	Y-Hex2HexA3HexNAc3Pen1S3-2H	3.02	462.3374	0
Y09	3	590.1316	1.08	Y-Hex2HexA3HexNAc3Pen1S2-2H	3.03	590.1334	1
Y10	4	506.3442	1.24	Y-Hex2HexA4HexNAc3Pen1S3-2H	2.47	506.3454	0

Y11	4	577.1052	4.18	Y-Hex2HexA4HexNAc4Pen1S4-2H	-1.26	577.1045	0
Y11	5	461.4807	0.88	Y-Hex2HexA4HexNAc4Pen1S4-2H	3.06	461.4821	0
Y13	5	553.2942	1.46	Y-Hex2HexA5HexNAc5Pen1S5-2H	2.89	553.2958	0
Y13	5	537.3031	2.20	Y-Hex2HexA5HexNAc5Pen1S4-2H	2.42	537.3044	1
Y15	5	645.1114	0.82	Y-Hex2HexA6HexNAc6Pen1S6-2H	-2.98	645.1095	0
Y15	5	629.1177	1.33	Y-Hex2HexA6HexNAc6Pen1S5-2H	0.70	629.1181	1
Y15	6	524.0957	0.70	Y-Hex2HexA6HexNAc6Pen1S5-2H	2.88	524.0972	1
Y17	6	613.9368	0.78	Y-Hex2HexA7HexNAc7Pen1S7-2H	-3.43	613.9347	0
Y17	7	526.0871	2.68	Y-Hex2HexA7HexNAc7Pen1S7-2H	-2.29	526.0859	0
Y19	7	591.6681	0.83	Y-Hex2HexA8HexNAc8Pen1S8-2H	-1.72	591.6670	0
C02	1	476.0702	5.52	HexA1HexNAc1S1-C	2.81	476.0715	0
C04	2	467.0649	1.34	HexA2HexNAc2S2-C	2.88	467.0663	0
C06	3	464.0635	0.84	HexA3HexNAc3S3-C	2.21	464.0645	0
C06	2	656.6216	0.73	HexA3HexNAc3S2-C	0.69	656.6220	1
INTERNAL CLEAVAGE	**Charge**	**M/Z**	**Relative Intensity**	**Ion**	**Mass Error (PPM)**	**M/Z (Theor.)**	
	2	559.5994	7.20	HexA2GalNAc3S2	2.32	559.6007	
	2	370.0439	7.15	HexA1GalNAc2S2	2.97	370.0450	
	1	300.0407	0.59	GalNAc1S1-O	-4.00	300.0395	
	1	282.0283	3.42	GalNAc1S1	2.13	282.0289	
	1	661.1392	5.94	HexA1GalNAc2S1	1.82	661.1404	

Supplementary Table 4.23. Assignment of fragment ions resulted from CID of parent ion m/z 570.2252 (z=14) corresponding to composition dp36-16S (f51). CID mass spectra were acquired on the Bruker Apex 9.4T FT-ICR mass spectrometer

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

5.1 CONCLUDING REMARKS

Glycosaminoglycans of varying lengths and numbers of modification can be characterized using this software. The non-database driven platform allowing for *in-silico* fragment development is both rapid, due to the repeating polymeric backbone of GAGs, and easily customizable for specific reducing end masses commonly observed in GAG-MS^2 experiments. Within the past 10-15 years, GAG work has been largely focused on disaccharide or tetrassacharides with the occasional glance at larger chains. This software has been designed with larger chains in mind, using a hexasaccharide to establish the mathematical system for fragment scoring and has also been applied to large (20-40 saccharide units) GAGs with moderate levels of sulfation. There is no theoretical limit on the GAG chain length that can be investigated with this method thus far, and continued refinement of the scoring model can only lead to further improvements.

5.2 FUTURE DIRECTIONS

Automated data interpretation and structural characterization is a challenge that is heavily reliant on creating code that in some facet replicates the expertise of individuals who are proficient at manual interpretation of mass spectra. The end goal is to put not only software but an entire laboratory workflow into the hands of clinicians and biologists and

other non-analytical chemists in order to assist them in answer fundamental biological questions. Software-based structural interpretation is often the final steps to large scale mass spectrometry related assays and can often make or break the end result. It is important to note that the software presented here is a comprehensive series of modules that is designed to focus on common elements typically observed in glycan mass spectrometry but is by not means immune to improvements or refinements. As glycomics in its entirety begins to be explored more, it becomes paramount to investigate mass spectral features that could be diagnostic for certain structural characteristics that are specific to a method or instrument. This information can and should be reintegrated into the scoring characteristics of the software to continue to refine the criteria. Innovation does not come without iteration, and iteration will be key to perfecting this software suite.