Software for Peak Finding and Elemental Composition Assignment for Glycosaminoglycan Tandem Mass Spectra*[S]

John D. Hogan‡§, Joshua A. Klein‡§, Jiandong Wu§, Pradeep Chopra¶, Geert-Jan Boons¶, Luis Carvalho‡‖, Cheng Lin§, and Joseph Zaia‡§**

Glycosaminoglycans (GAGs) covalently linked to proteoglycans (PGs) are characterized by repeating disaccharide units and variable sulfation patterns along the chain. GAG length and sulfation patterns impact disease etiology, cellular signaling, and structural support for cells. We and others have demonstrated the usefulness of tandem mass spectrometry (MS^2) for assigning the structures of GAG saccharides; however, manual interpretation of tandem mass spectra is time-consuming, so computational methods must be employed. In the proteomics domain, the identification of monoisotopic peaks and charge states relies on algorithms that use averagine, or the average building block of the compound class being analyzed. Although these methods perform well for protein and peptide spectra, they perform poorly on GAG tandem mass spectra, because a single average building block does not characterize the variable sulfation of GAG disaccharide units. In addition, it is necessary to assign product ion isotope patterns to interpret the tandem mass spectra of GAG saccharides. To address these problems, we developed GAGfinder, the first tandem mass spectrum peak finding algorithm developed specifically for GAGs. We define peak finding as assigning experimental isotopic peaks directly to a given product ion composition, as opposed to deconvolution or peak picking, which are terms more accurately describing the existing methods previously mentioned. GAGfinder is a targeted, brute force approach to spectrum analysis that uses precursor composition information to generate all theoretical fragments. GAGfinder also performs peak isotope composition annotation, which is typically a subsequent step for averagine-based methods. Data are available via ProteomeXchange with identifier PXD009101. ***Molecular & Cellular Proteomics 17: 1448–1456, 2018. DOI: 10.1074/mcp.RA118.000590.***

Glycosaminoglycans (GAGs)[1] exist either as the glycan portion of proteoglycans (PGs) or as extracellular matrix (ECM) polysaccharides. The three classes of sulfated GAGs, heparan sulfate (HS), chondroitin sulfate (CS), and keratan sulfate (KS), are characterized by their long, linear chain, a repeating disaccharide unit (specific to each GAG class), and variable patterns of sulfation and acetylation. Because of their locations on the cell surface and in the ECM, as well as their sequence variation, they interact with many growth factors and growth factor receptors and therefore modulate cellular signaling and signal transduction pathways (1–2). Furthermore, spatial and temporal regulation of the structures of GAGs characterizes physiology and pathophysiology in eukaryotes. For instance, cancer cells remodel HS chains in their microenvironments to avoid immune system targeting and allow proliferation (3). In the motor neuron-degenerative disease amyotrophic lateral sclerosis, KS sulfation has been shown to correlate with disease progression (4). Indeed, GAG expression is required for embryonic development (5), and GAGs are required for the proper functioning of all mammalian biological systems (1). Clearly, assigning GAG sequences from tandem mass spectral data is necessary to establish their roles in diverse disease mechanisms.

Tandem mass spectrometry (MS^2) entails isolating a precursor ion in the first stage and dissociating it in subsequent stages. Manual interpretation of tandem mass spectra is tedious, time-consuming, and subjective. The first step of interpretation is to assign the *m/z* and charge states for product ions. Once this is done, neutral masses and isotope compositions can be assigned. Once these assignments are made, an algorithm can be used to identify the GAG sequence (7).

Wolff and colleagues first applied electron activated dissociation methods to GAG oligosaccharides, using both electron detachment dissociation (EDD) (8) and negative electron transfer dissociation (NETD) (9). More recently, Huang and colleagues showed the effectiveness of electron activated dissociation for minimizing sulfate loss during HS mass spectrometry experiments (10). Resulting tandem mass spectra after electron activated dissociation are extremely rich in that

From the ‡Program in Bioinformatics, Boston University – Boston, MA 02215; §Center for Biomedical Mass Spectrometry, Department of Biochemistry, Boston University School of Medicine – Boston, MA 02118; ¶Complex Carbohydrate Research Center, University of Georgia – Athens, GA 30602; ‖Department of Mathematics & Statistics, Boston University – Boston, MA 02215

Received January 6, 2018, and in revised form, March 25, 2018
Published, MCP Papers in Press, April 3, 2018, DOI 10.1074/mcp.RA118.000590

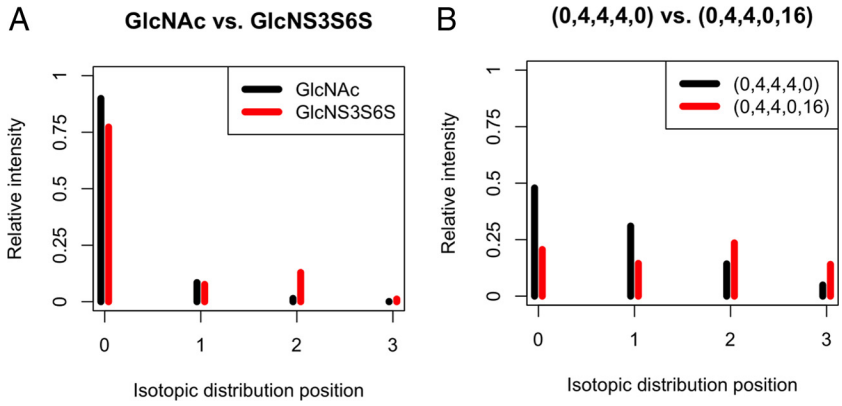

FIG. 1. **Comparison of expected isotopic distributions for oligosaccharides with varying sulfation.** *A*, Expected isotopic distribution of non-sulfated *N*-acetylglucosamine (GlcNAc) compared with 3,6-*O*-sulfated, *N*-sulfated glucosamine (GlcNS3S6S). *B*, Expected isotopic distribution of a non-sulfated octasaccharide with acetyl groups at all four *N* positions compared with a hexadecasulfated octasaccharide with sulfate groups at every possible position. Notice the higher intensity at the A+2 peak for each fully sulfated oligosaccharide; for the octasaccharide, the A+2 peak has the highest intensity, making monoisotopic peak detection more difficult. Intensity is relative to the total intensity for the whole isotopic distribution. Key for octasaccharide: [ΔHexA, HexA, GlcN, Ac, SO_3].

they contain many product ions with varying charge states and isotope patterns. In the proteomics domain, several computational methods for automatic recognition of isotopic patterns and assignment of charge states and neutral mass values have been developed, including THRASH (11), Decon2LS (12), and MS-Deconv (13), among others. These methods assume product ion isotopic distributions will match the pattern produced by the molecule's average building block, or averagine; however, performance for GAG saccharide tandem mass spectra is inadequate, because of the variable levels of sulfation along their chains and the relatively abundant ^{34}S isotope. Fig. 1 shows two examples of the large difference in the expected isotopic distributions of non-sulfated and fully sulfated GAG fragments. Plainly, there is no GAG averagine that would accurately recover the correct monoisotopic peak for each fragment, and that leads to incorrect and missing assignments. Averagine-based approaches also do not assign elemental compositions for monoisotopic ions, a step necessary for interpretation of GAG saccharide tandem mass spectra. We sought to solve these problems.

Previous work in GAG tandem mass spectra analysis and annotation has typically been a step in a further sequencing project. For instance, Yu and colleagues recently sequenced the dermatan sulfate (DS) chain of the pericellular PG decorin using a genetic algorithm based on known sulfate modification information from disaccharide analysis but mentioned in-house data interpretation software in passing (14). And two GAG sequencing efforts from Chiu and colleagues, GAG-ID (15) and a multivariate mixture model to estimate identification accuracy (16) represent recent attempts at automated GAG sequencing using a weighted hypergeometric distribution to match spectra to potential sequences. However, these papers both describe a method that only considers high intensity peaks, rather than full isotopic distributions, and their method requires an intense experimental workup for chemical derivatization that replaces sulfate groups with heavy isotope acetyl groups.

Averagine-based deisotoping and charge state deconvolution algorithms were developed to circumvent the combinatorial explosion of the number of possible protein sequences as the length of the chain increases. Because of this expansion, brute force methods searching all possible proteins and protein product ions are not feasible. Although the number of possible GAGs also increases exponentially as a function of chain length, the rate of increase is much lower. Fig. 2 shows the $\log_{10}$ of the number of possible structures of unmodified proteins, HS GAG saccharides, CS GAG saccharides, and KS GAG saccharides, as a function of the length of the chain. Notice how the slopes for each GAG class are much smaller than the slope for proteins and consider how many more protein structures are possible when post-translational modifications are included. Given the reduced search space and the variable sulfation along GAG chains, we developed a brute force product ion search algorithm using the Python programming language, GAGfinder, for MS^2 of GAG saccharides of a given composition. GAGfinder iterates through every possible fragment of a GAG composition at multiple charge states and tests its theoretical isotopic distribution against the observed spectral pattern. GAGfinder is available for download at http://www.bumc.bu.edu/msr/software. This paper describes the steps in GAGfinder and its performance as a means to identify the GAG monoisotopic product ions, charge states, and neu-

[1] The abbreviations used are: GAG, Glycosaminoglycan; AUC, Area under the curve; CS, Chondroitin sulfate; ECM, Extracellular matrix; EDD, Electron detachment dissociation; HS, Heparan sulfate; KS, Keratan sulfate; MS^2, Tandem mass spectrometry; NETD, Negative electron transfer dissociation; PG, Proteoglycan; ppm, Parts-per-million; S/N, Signal-to-noise ratio; TIC, Total ion current.

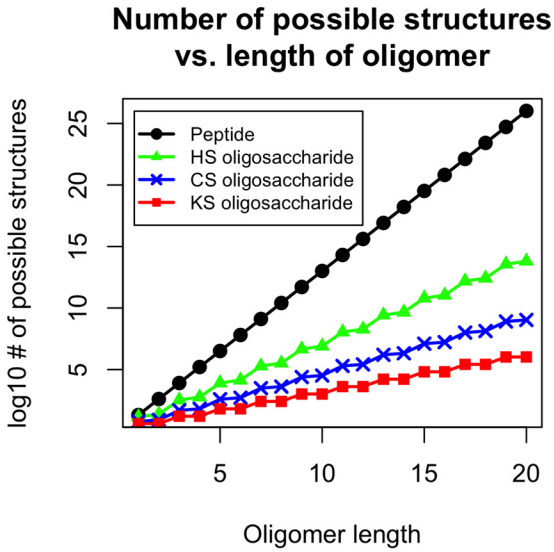

FIG. 2. **Plot of $\log_{10}$ of the number of possible structures given oligomer length.** The number of unmodified protein sequences of a given oligomer length grows at a much faster rate than those of HS, CS, or KS. The slower combinatorial growth rate allows GAGfinder's brute force search to be feasible.

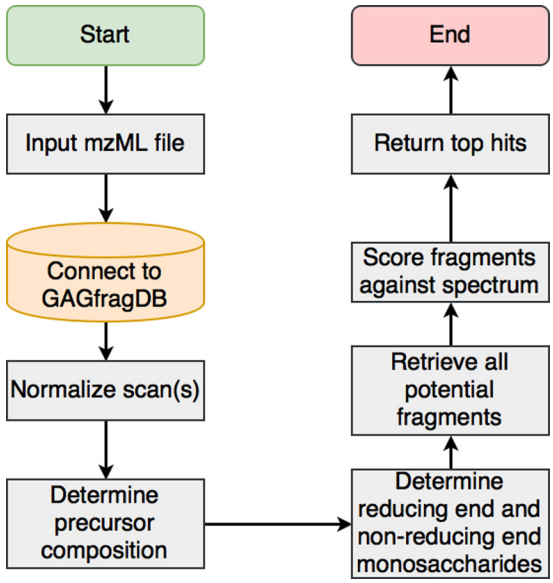

FIG. 3. **Workflow for GAGfinder.** The steps in GAGfinder's algorithm.

tral mass values *versus* an averagine-based peak finding algorithm.

EXPERIMENTAL PROCEDURES

GAGfinder Overview—A flowchart of the steps GAGfinder can be viewed in Fig. 3. The details of each step are described below. The term "product ion" will be used to refer to ions observed in tandem mass spectra. The term "fragment" will be used to refer to theoretical GAG saccharide substructures in a database.

Inputs—There are several required and optional inputs for GAGfinder to return accurate results. The spectrum data must be in the mzML file format (17); the raw data can be converted using any format conversion tool, such as MSConvert (18) or compassXport (Bruker Daltonics, Inc.). Other required inputs include the GAG class, the precursor *m/z*, the precursor charge, and the output format for the results. Either the top percentile or the top N results can be returned, but not both. Optional inputs include the reducing-end derivatization formula (if any), the adducted metal and the number of adducts (if there is metal adduction), the NETD cation reagent (if NETD), a user-specified internal precision for mapping fragments to isotopic distributions, a Boolean value for whether noise has already been removed from the spectrum, and the number of labile sulfate losses to consider. These inputs are arguments for the GAGfinder command line program.

Step 1: Load mzML File and Connect to GAG Fragment Database—The first step of GAGfinder is connecting to GAGfragDB, the database developed in SQLite for easy storing and retrieval of all possible fragments of a precursor composition up to hexadecamer. There are 4150 unique compositions, 65,664 fragments, and 17,156,928 precursor-fragment mappings in GAGfragDB. The composition with the most possible fragments - (1, 7, 8, 4, 15) with a key of (dHexA, HexA, HexN, Ac, SO_3) - has 21,299 child fragments associated with it in HS. GAGfragDB includes a controlled vocabulary designed to give each fragment a unique text identifier that does not assume anything about the structure of the precursor or the fragment. In other words, a fragment that has one composition but could be a terminal fragment or any number of internal fragments will have only one identifier. Supplemental Fig. S1 shows the relational schema for GAGfragDB. The connection to GAGfragDB is established by the Python sqlite3 module. After connecting to GAGfragDB, GAGfinder loads the mzML file into Python using the pymzML module (19). The pymzML module has several spectrum processing methods, including centroiding peaks, finding peaks in the spectrum within a particular error tolerance, and a number of others.

Step 2: Normalize Scan(s) and Remove Noise—Once the tandem mass spectral data have been loaded into Python, GAGfinder normalizes and averages the scans of the data file using the total ion current (TIC). GAGfinder first divides each scan in the file by the summed TIC intensity and then calculates the average over all scans. This step prevents any of the scans from biasing the results over the rest of the scans and is performed using methods in the pymzML package. After normalizing the scans, GAGfinder removes noise from the spectrum, if the spectrum has not already been denoised by the user prior to runtime. GAGfinder uses an implementation of the noise reduction algorithm MasSPIKE (20).

Step 3: Determine Precursor Composition—Given the precursor *m/z* and charge, the neutral mass of the precursor can be calculated, and based on this and the GAG class, the precursor composition can be determined. GAGfinder considers metal adduction and reducing end derivatization information to calculate the neutral mass matching the composition in GAGfragDB. GAGfinder selects the composition with the neutral mass closest to the calculated precursor mass as the precursor composition.

Step 4: Determine Reducing End and Nonreducing End Monosaccharides—In order to reduce the search space as much as possible, GAGfinder attempts to determine the monosaccharides at each precursor saccharide terminus. There are several cases in which this is possible, and Fig. 4, shows the decision tree for determining this. First, if the non-reducing end is an unsaturated uronic acid (in the cases of CS and HS saccharides generated by polysaccharide lyase enzyme digestion), GAGfinder first assumes that the reducing end monosaccharide is a hexuronic acid if the precursor contains an odd number of monosaccharides, and a hexosamine if the precursor contains an even number of monosaccharides. If this is not the case, then GAGfinder checks whether there is an unequal number of the parts of the repeating disaccharide for the current GAG class. If the

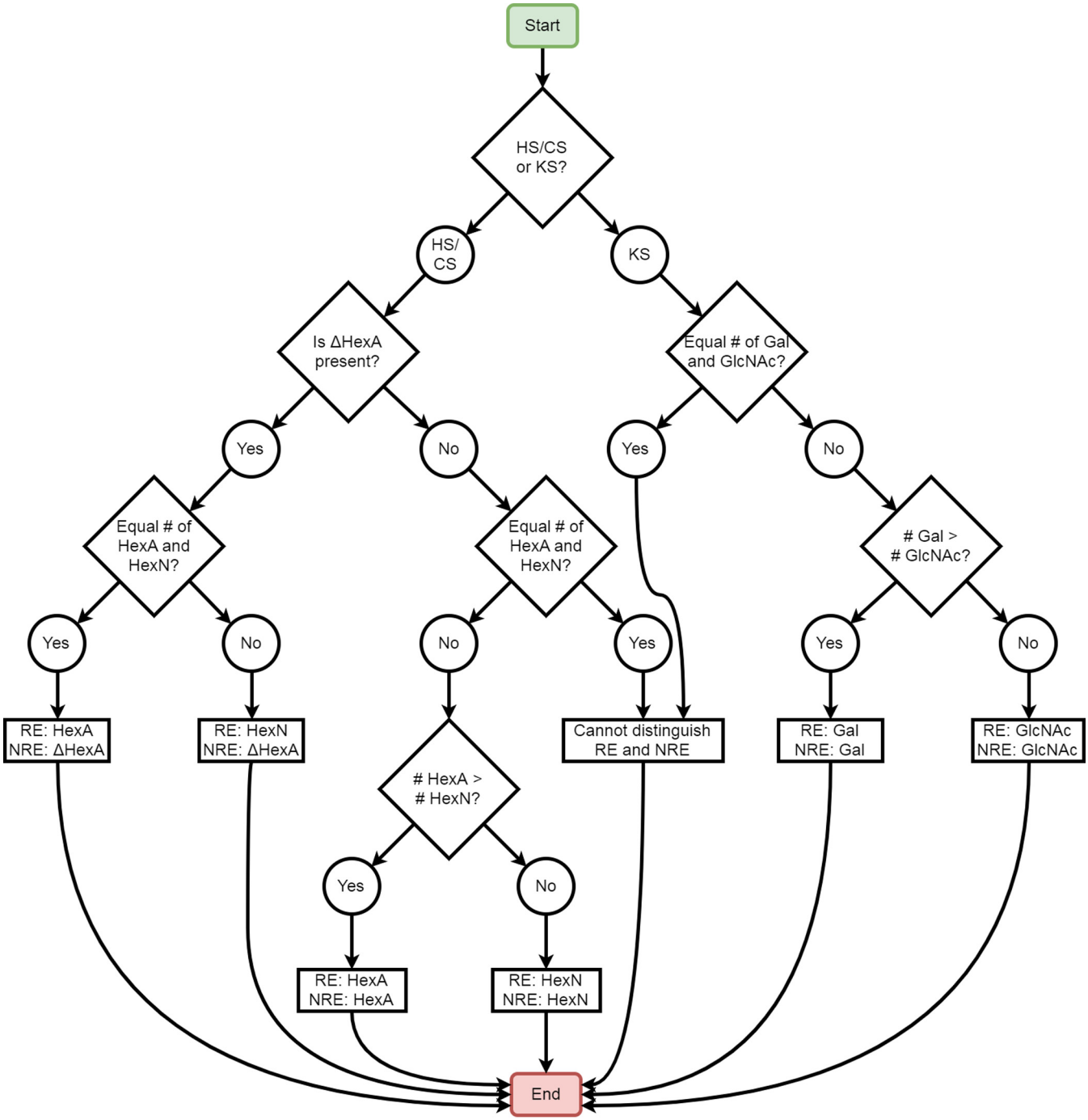

FIG. 4. **Flowchart describing steps in determining terminal sugars.** In several cases, GAGfinder can determine the reducing and non-reducing end sugars based on biosynthetic rules. In cases where the sugars cannot be distinguished from the composition, both monosaccharides of the class of GAG are considered as the terminal sugars. RE = reducing end; NRE = non-reducing end.

number is unequal, then whichever monosaccharide there is more of will be on both the nonreducing and reducing end. If the number is equal, then GAGfinder cannot assign the end fragments and must search through the entire search space.

Step 5: Retrieve and Modify All Theoretical Fragments for the Precursor—Next, GAGfinder retrieves every possible fragment for the current precursor from GAGfragDB. The possible fragments stored in GAGfragDB include glycosidic bond cleavages and all cross-ring cleavages except for those involving cleavage of adjacent bonds. Supplemental Fig. S2 shows each cross-ring cleavage GAGfinder considers. GAGfragDB stores the theoretical fragments as neutral masses without considering sulfate losses or any other modification information, so GAGfinder must modify and search each fragment in order to maximize spectrum coverage. For each fragment, the modifications included are water loss (for glycosidic fragments only), hydrogen loss (up to 2), sulfate loss (up to the amount designated by

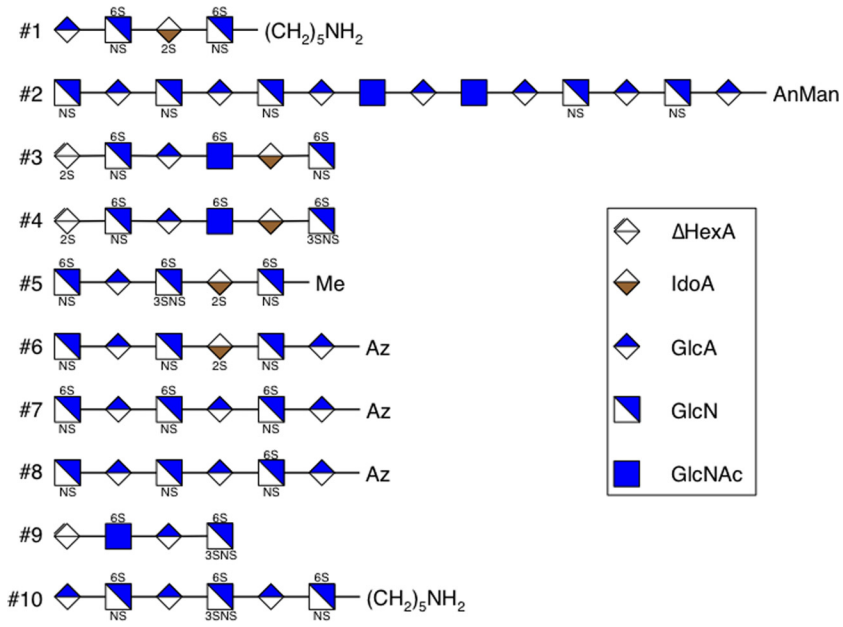

FIG. 5. **Structures of the ten synthetic standards used for testing purposes.** #1 has charge state of 4- and dissociation method of NETD. #2 has charge state of 8- and dissociation method of NETD. #3 has charge state of 5- and dissociation method of NETD. #4 has charge state of 4- and dissociation method of EDD. #5 has charge state of 6- and dissociation method of EDD. #6 has charge state of 4- and dissociation method of NETD. #7 has charge state of 4- and dissociation method of NETD. #8 has charge state of 3- and dissociation method of EDD. #9 has charge state of 3- and dissociation method of NETD. #10 has charge state of 4- and dissociation method of EDD. These standards were selected randomly because of their range of modifications, length, and different dissociation methods.

the user) and reducing end derivatization (if any). This information is used to determine whether a given fragment corresponds to the reducing terminus. Product ions that have the same chemical composition are merged. For every combination of these modifications, the fragments are pushed through the algorithm.

Step 6: Score Each Theoretical Fragment—Once all the theoretical fragments have been retrieved and modified as need be, they are scored against the tandem mass spectrum. GAGfinder considers charge states from −1 to that of the precursor ion plus one for each fragment. The decision to use the charge state of the precursor ion plus one for the upper bound rather than that of the precursor ion is because of two main reasons. First, the number of product ions with the same charge state as the precursor is a small percentage of all of the product ions, meaning including this charge state in GAGfinder's searching would find only a few more product ions while introducing more false positives. Second, many of the product ions with the same charge state as the precursor are derivatives of the precursor, meaning they provide no additional structural information. A theoretical relative isotopic distribution (TID) is calculated for each fragment using the BRAIN algorithm (21), which employs polynomial expansion and applies the Newton-Girard theorem and Viète's formulae to this end. Once the TID is calculated, GAGfinder searches the tandem mass spectrum for product ion peaks at the *m/z* values of the TID within either a user-specified error tolerance or the default error tolerance of 20 parts-per-million (ppm), storing them as the experimental isotopic distribution (EID). The EID is then divided by the sum of its intensities so that it is also a relative distribution. GAGfinder employs a G-test of goodness-of-fit to determine how similar the EID is to the TID. Equation 1 shows the expression for the G score, where *i* is the index of each peak in the matched isotopic distributions. According to the G-test, the G score follows a chi-squared distribution under the null hypothesis that the EID has the same distribution as the TID, and so can be used to compute *p* values. This way, a lower G score yields a higher *p* value and thus represents a better fit.

$$G = 2\sum EID_i \ln\left(\frac{EID_i}{TID_i}\right) \qquad \text{(Eq. 1)}$$

Step 7: Rank Product Ions by G Score and Return Top Hits—Once all theoretical fragments have been scored for goodness-of-fit, they are ranked by increasing G score. Depending on whether the user requested the top percentile or top N results, those results are saved into an output file. The output file contains the fragment *m/z*, charge, intensity, annotation(s), G score, and error in ppm.

Data Acquisition and Preprocessing—We chose ten synthetic GAG standards to demonstrate the effectiveness of GAGfinder (Fig. 5). These standards were chosen because of their range of modification distribution and precursor charges. Compounds 1 and 10 were synthesized as described (22). Compound 2 was a generous gift from Prof. Jian Liu, University of North Carolina, Chapel Hill. Compound 3 was purchased from New England Biolabs (Andover, MA). Compound 5 was purchased as Arixtra pharmaceutical preparation and desalted by size exclusion chromatography. Compounds 4, 6, 7, and 8 were acquired through a publicly available set of HS standard saccharides funded by the NIH and maintained by the Zaia laboratory (http://www.bumc.bu.edu/zaia/gag-synthetic-saccharides-available/). Compound 9 was isolated from porcine intestinal mucosa as described (23). These were subjected to electron detachment dissociation (EDD) or negative electron transfer dissociation (NETD) using a Bruker solariX 12T FTMS instrument. For each saccharide, GAGfinder was run retrieving 100% of tested fragments, allowing for two sulfate losses, and using the default error of 20 ppm when mapping fragments to isotopic distributions. For saccharides 1–5, noise was not previously removed, so GAGfinder implemented MasSPIKE to remove noise. For saccharides 6–10, noise was previously removed. Although in principle GAGfinder can handle all classes of GAGs, we show results for HS saccharides for the present work. Details regarding the tandem mass spectrometric acquisition methods can be found in Hu *et al.* (7). Raw data files were converted to mzML format

for input into GAGfinder by either MSConvert GUI version 3.0.5084 (13) or compassXport command line utility 3.0.13 (Bruker Daltonics, Inc.). The mass spectrometry glycomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (24) partner repository with the data set identifier PXD009101.

We first sought to demonstrate the ability of GAGfinder to identify product ion isotope clusters and charge states. To do this, we generated a list of product ions using a traditional averagine-based method (the SNAP peak finder in Bruker DataAnalysis 4.2) *versus* that for GAGfinder. In order to retrieve every product ion SNAP identified, we set the quality factor threshold at 0, the signal-to-noise ratio (S/N) threshold at 1, the relative intensity threshold (base peak) at 0%, and the absolute intensity threshold at 0. For each GAG saccharide tested, we set the maximum charge state to the absolute value of the precursor charge state minus one, so that SNAP would behave comparatively to GAGfinder. We set the repetitive building block to $C_6H_{11.375}N_{1.125}O_{9.5}S_{1.5}$, as used in previous methods (25). SNAP returned a matrix with columns for *m/z*, charge, intensity, resolving power, and quality factor.

Method Comparison—In order to judge GAGfinder's performance in assigning tandem mass spectral monoisotopic product ions and charge states, we employed two separate statistical methods. Each method required unbiased expert manual selection of monoisotopic product ion peaks to serve as the set of true positives. In both methods we had GAGfinder return scores for 100% of the tested theoretical fragments to ensure maximum spectral coverage. The first method compared the GAGfinder performance against that of a random selection of monoisotopic product ions. The second compared GAGfinder's performance to that of an averagine-based peak finding algorithm.

The first method for judging GAGfinder's performance was a permutation test that gauged GAGfinder's performance in selecting true positive product ion peaks compared against random selection of product ion peaks. First, we calculated a performance score (PerfScore) for the GAGfinder results using the equation

$$PerfScore = \sum_j G_j\, Hit_j \qquad \text{(Eq. 2)}$$

where j is the index of the current product ion, G_j is the G score for fragment j, and

$$Hit_j = \begin{cases} 1, \text{ if product i on j is a "real" hit} \\ 0, \text{ if product i on j is not a "real" hit} \end{cases} \qquad \text{(Eq. 3)}$$

Once we calculated the performance score for the GAGfinder results, we permuted the *Hit* vector 10,000 times and recalculated the performance score for each permutation. Because G scores are smaller for better fits, a smaller performance score represents a better performance. The performance scores of the 10,000 permutations represent a background distribution for performance against which we compared the GAGfinder performance score. We plotted GAGfinder's performance score against the background distribution and recorded its rank among all the permuted performance scores.

The second method for testing GAGfinder's performance was a binary classifier evaluation that compared the GAGfinder performance *versus* that of an averagine based algorithm, SNAP. Precision-recall (P-R) curves show how the classifier's precision and recall change as the classifier's threshold is changed, and the area under the curve (AUC) represents the classifier's performance. Precision is defined as

$$Precision = \frac{TP}{TP + FP} \qquad \text{(Eq. 4)}$$

where TP stands for true positives and FP stands for false positives, and recall, also known as sensitivity, is defined as

$$Recall = \frac{TP}{TP + FN} \qquad \text{(Eq. 5)}$$

where TP stands for true positives and FN stands for false negatives. A perfect classifier has a precision and a recall of 1, and therefore, the closer the P-R AUC is to 1, the better the classifier has performed. The complete results for GAGfinder were generated by requesting 100% of the product ions tested.

For GAGfinder results, we generated the vector of precision and recall values by ordering the results by G-score in ascending order and calculating the precision and recall of GAGfinder at each G-score threshold. Similarly, for SNAP, we ordered the results by quality factor in descending order and calculated its precision and recall at each quality factor threshold. Because the number of true positive peaks is limited to those fragment masses extracted from GAGfragDB, GAGfinder identifies fewer monoisotopic peaks and charge states than does an averagine-based algorithm. To compare the effectiveness of the peaks assigned in common by both algorithms, we removed peaks that were not searched by GAGfinder. These peaks were likely because of fragmentation or chemistry that GAGfinder does not consider.

RESULTS

GAGfinder Performance Compared with Random Sampling—For each of the ten GAG saccharide tandem mass spectra tested, the GAGfinder performance score significantly outperformed that of the permutations. Table I compares GAGfinder's performance score *versus* the mean and standard deviation of the 10,000 random permutations for each saccharide; the distribution plots for each saccharide can be seen in supplemental Fig. S3. For every compound, the PerfScore for GAGfinder was in the top ten lowest scores, and for seven of the compounds, the PerfScore for GAGfinder was lower than every permutation's PerfScore. This indicated that GAGfinder produced a better performance than a random selection. Furthermore, the GAGfinder PerfScore was at least three standard deviations lower than the average of the permutations for every saccharide, signifying significant outperformance compared with a random selection of peaks. There was no correlation between the PerfScores for GAGfinder and the means and standard deviations of the permutations, the dissociation method, or the precursor charge state. This indicated a lack of bias for the GAGfinder algorithm. We concluded based on these numbers that GAGfinder significantly outperforms a random selection of peaks.

GAGfinder Performance Compared with Averagine-based Peak Finding—Table II compares the P-R curve AUCs for each spectrum for GAGfinder *versus* SNAP, including summary statistics. The P-R curves for each spectrum are shown in supplemental Fig. S4. The average GAGfinder P-R AUC is higher than the average SNAP P-R AUC, whereas the median GAGfinder P-R AUC is almost equal to the median SNAP P-R AUC. For seven of the ten spectra, GAGfinder has a higher P-R AUC. Of these seven, four were generated by NETD, whereas the other three were generated by EDD, and there is no correlation between charge state and performance difference between GAGfinder and SNAP, indicating a lack of bias in the performance of each. These numbers show that GAG-

TABLE I
Performance scores for GAGfinder compared to the mean and standard deviation of the 10,000 permutations for each of the ten synthetic compounds

In each case, GAGfinder's PerfScore was lower than at least 99.9% of the permutations', indicating a better performance.

Compound	Precursor z	Dissociation method	GAGfinder PerfScore	Permutation mean PerfScore	Permutation PerfScore std. dev.	Rank
#1	4-	NETD	32.502	42.072	2.983	2/10,001
#2	8-	NETD	60.866	112.712	6.176	1/10,001
#3	5-	NETD	63.966	83.238	5.487	3/10,001
#4	4-	EDD	58.422	89.935	6.000	1/10,001
#5	6-	EDD	60.991	74.966	4.583	9/10,001
#6	4-	NETD	58.492	103.262	7.304	1/10,001
#7	4-	NETD	37.769	89.721	6.411	1/10,001
#8	3-	EDD	31.853	65.747	4.741	1/10,001
#9	3-	NETD	14.390	40.784	4.904	1/10,001
#10	4-	EDD	66.811	118.989	7.333	1/10,001

TABLE II
Area under the curve (AUC) of precision-recall (PR) curves for GAGfinder analysis results compared to those from SNAP

Compound	Precursor z	Dissociation method	GAGfinder AUC	SNAP AUC
#1	4-	NETD	0.681	0.765
#2	8-	NETD	0.315	0.100
#3	5-	NETD	0.612	0.229
#4	4-	EDD	0.688	0.652
#5	6-	EDD	0.611	0.536
#6	4-	NETD	0.628	0.574
#7	4-	NETD	0.635	0.735
#8	3-	EDD	0.716	0.682
#9	3-	NETD	0.792	0.619
#10	4-	EDD	0.646	0.703
		Mean	0.632	0.560
		Std. Dev.	0.125	0.222
		Median	0.641	0.636

finder identifies monoisotopic peaks and charge states with similar accuracy as does the averagine-based SNAP algorithm. We note again that GAGfinder assigns the elemental compositions for all identified monoisotopic peaks.

Runtime Numbers for GAGfinder—GAGfinder tracks and reports the length of runtime for each analysis. The amount of time required for GAGfinder to search each fragment varies based on a variety of factors, but the two that affect runtime the most are the number of possible fragments and whether the noise was removed prior to analysis. Table III shows GAGfinder's runtime for each saccharide, as well as the total number of fragments for that composition and charge state combination and whether or not the data was pre-processed. As can be seen, analyzing a spectrum without noise removed greatly increases the runtime. This is not because of GAGfinder's noise removal step taking an inordinate amount of time, but rather because of the larger number of data points to average across scans. For instance, samples 1–5 did not have noise removed prior to analysis, leaving that step for GAGfinder, which slowed down runtime. However, samples 6–10 did have noise removed prior to analysis, and their faster runtime shows it.

DISCUSSION

Here we have presented GAGfinder, the first GAG-specific isotopic distribution finding software for high resolution tandem mass spectra. GAGfinder uses a targeted, brute force approach to search observed product ions against a set of theoretical fragments calculated based on the precursor ion exact mass, composition based on GAG biosynthesis rules, and expected NETD and EDD tandem mass spectrometry dissociation patterns. The software is easy to use on any operating system and outperforms traditional peak finding software that was designed for peptide fragments. For this manuscript, GAGfinder was run as a command line utility on a MacBook Pro, and all tandem mass spectrometric data are available on the PRIDE Proteomics IDEntifications archive. Although the software is currently only available in command line form, a web application and interface is currently under development and will be available soon.

We tested GAGfinder on the EDD and NETD spectra of a diverse set of synthetic GAGs and showed that it accurately and consistently returns valid fragments for the precursor being tested. GAGfinder consistently scored true positive

TABLE III
Runtime for GAGfinder analysis for each saccharide

Compound	Precursor z	Dissociation method	Runtime (s)	# of tested fragments/# of possible fragments	Noise removed?
#1	4-	NETD	124.450	122/4,089	No
#2	8-	NETD	176.900	827/99,120	No
#3	5-	NETD	139.566	177/12,256	No
#4	4-	EDD	76.679	252/9,147	No
#5	6-	EDD	71.833	172/6,870	No
#6	4-	NETD	1.740	413/7,398	Yes
#7	4-	NETD	1.885	391/8,373	Yes
#8	3-	EDD	1.724	257/4,932	Yes
#9	3-	NETD	1.653	160/2,524	Yes
#10	4-	EDD	1.830	428/8,379	Yes

fragments better than false fragments across all tested GAGs and performed comparably to traditional peak finding methods. Unlike traditional peak finding methods, GAGfinder assigns elemental compositions to the monoisotopic product ions that are essential for assigning the saccharide structure. Although we tested GAGfinder exclusively on high resolution spectra in the negative ion mode, the software was designed in principle to handle any resolution level in either the negative or positive ion mode. For low resolution spectra, we hypothesize that the G-scores for assigned monoisotopic product ions will be worse than with high resolution data; however, this is because of the whole distribution of G-scores shifting, and we anticipate that the correct IDs will still be found at or near the top of the ranked list of G-scores.

Although GAGfinder succeeds at identifying product ions that fall within the set defined in the GAGfragDB, it does not identify product ions that arise from undefined dissociation processes. Such undefined processes include rare dissociation patterns, a charge state equal to or higher in absolute value than that of the precursor, and random instrument noise. In these cases, traditional methods will have a greater likelihood of identifying *m/z* values and charge states but will not identify the elemental composition. Furthermore, these ions are not actually useful for GAG structure determination, which is the goal of GAG sequencing. Although it is possible to add rare dissociation processes to the GAGfragDB, this would increase search space size at the expense of algorithm run time.

An interesting case where GAGfinder outperforms the traditional peak finding method SNAP arises when the fragment composition substantially differs from that of the averagine used. As shown in Fig. 1, selecting an appropriate averagine that fits all GAG fragments is difficult because of the variable number of sulfur atoms in the fragments. Compound #9 contains a heavily sulfated reducing end, with three sulfate groups on one GlcNAc. Although GAGfinder finds the Y1-S and Y1 ions for this compound and scores them in the top ten, SNAP is unable to find them. supplemental Fig. S5 shows the annotated spectra, using the top 20 (or so) most intense fragments for each saccharide, and supplemental Fig. S6 shows the portion of the spectrum containing these fragments. In both cases, there are other isotopic distributions interspersed, but none of these precisely overlap with their peaks.

Wolff and colleagues first showed how metal cationization can help curb sulfate loss in EDD (6), an approach that has gained popularity in the years since. Although our group typically avoids metal adduction during GAG analysis because of the negative effects on the instrument and the extra work up, we nonetheless designed GAGfinder to be able to handle samples that have been cationized. In GAGfinder, cationization adds to the search space, and therefore the runtime, without necessarily improving peak finding performance, reduced sulfate loss aside. Although metal cationization can help remove the ambiguity of tandem mass spectra, allowing for easier GAG sequencing, its utility is seen mostly in that step of the sequencing pipeline. GAGfinder is only looking for fragments and isotopic distributions of given compositions, regardless of whether there is metal cationization or not, and therefore, metal cationization should not affect GAGfinder's peak finding performance.

In conclusion, use of GAGfinder will allow researchers to swiftly and accurately assign elemental compositions and product ion types to product ions in GAG saccharide tandem mass spectra. Although GAGfinder was tested exclusively on pure, synthetic compounds, we are evaluating its ability to assign product ion *m/z*, charge state, and elemental composition for biological samples. Finally, we demonstrate that the use of a brute force method for peak finding balances search space size and overall analysis time compared with traditional methods.

DATA AVAILABILITY

The raw mass spectrometry data are available at ProteomeXchange via the PRIDE database with the accession number PXD009101.

* This work was funded by NIH grants R21HL131554, U01CA221234 and P41GM104603.

[S] This article contains supplemental material.

** To whom correspondence should be addressed: Boston University Medical Campus, 670 Albany St., Rm. 509, Boston, MA 02118. Tel.: +1 (617) 638 6762; E-mail: jzaia@bu.edu.

Author contributions: J.D.H., J.A.K., and J.W. performed research; J.D.H. and J.W. analyzed data; J.D.H. wrote the paper; P.C. and G.-J.B. contributed new reagents/analytic tools; L.C., C.L., and J.Z. designed research.

REFERENCES

1. Bishop, J. R., Schuksz, M., and Esko, J. D. (2007) Heparan sulphate proteoglycans fine-tune mammalian physiology. *Nature* **446,** 1030–1037
2. Lindahl, U., and Li, J. P. (2009) Interactions between heparan sulfate and proteins-design and functional implications. *Int. Rev. Cell Mol. Biol.* **276,** 105–159
3. Fuster, M. M., and Esko, J. D. (2005) The sweet and sour of cancer: glycans as novel therapeutic targets. *Nat. Rev. Cancer* **5,** 526–542
4. Hirano, K., Ohgomori, T., Kobayashi, K., Tanaka, F., Matsumoto, T., Natori, T., Matsuyama, Y., Uchimura, K., Sakamoto, K., Takeuchi, H., Hirakawa, A., Suzumura, A., Sobue, G., Ishiguro, N., Imagama, S., and Kadomatsu, K. (2013) Ablation of Keratan Sulfate Accelerates Early Phase Pathogenesis of ALS. *PLoS ONE* **8,** e66969
5. Perrimon, N., and Bernfield, M. (2000) Specificities of heparan sulfate proteoglycans in developmental processes. *Nature* **404,** 725–728
6. Wolff, J. J., Laremore, T. N., Busch, A. M., Linhardt, R. J., and Amster, I. J. (2008) Influence of charge state and sodium cationization on the electron detachment dissociation and infrared multiphoton dissociation of glycosaminoglycan oligosaccharides. *J. Am. Soc. Mass Spectrom.* **19,** 790–798
7. Hu, H., Huang, Y., Mao, Y., Yu, X., Xu, Y., Liu, J., Zong, C., Boons, G., Lin, C., Xia, Y., and Zaia, J. (2014) A computational framework for heparan sulfate sequencing using high-resolution tandem mass spectra. *Mol. Cell. Proteomics* **13,** 2490–2502
8. Wolff, J. J., Chi, L., Linhardt, R. J., and Amster, I. J. (2007) Distinguishing glucuronic from iduronic acid in glycosaminoglycan tetrasaccharides by using electron detachment dissociation. *Anal. Chem.* **79,** 2015–2022
9. Wolff, J. J., Leach III, F. E., Laremore, T. N., Kaplan, D., Easterling, M. E., Linhardt, R. J., and Amster, I. J. (2010) Negative electron transfer dissociation of glycosaminoglycans. *Anal. Chem.* **82,** 3460–3466
10. Huang, Y., Yu, X., Mao, Y., Costello, C. E., Zaia, J., and Lin, C. (2013) De novo sequencing of heparan sulfate oligosaccharides by electron-activated dissociation. *Anal. Chem.* **85,** 11979–11986
11. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **11,** 320–332
12. Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinf.* **10,** 87
13. Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., and Pevzner, P. A. (2010) Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Mol. Cell. Proteomics* **9,** 2772–2782
14. Yu, Y., Duan, J., Leach III, F. E., Toida, T., Higashi, K., Zhang, H., Zhang, F., Amster, I. J., and Linhardt, R. J. (2017) Sequencing the dermatan sulfate chain of Decorin. *J. Am. Chem. Soc.* **139,** 16986–16995
15. Chiu, Y., Huang, R., Orlando, R., and Sharp, J. S. (2015) GAG-ID: Heparan sulfate (HS) and heparin glycosaminoglycan high-throughput identification software. *Mol. Cell. Proteomics* **14,** 1720–1730
16. Chiu, Y., Schliekelman, P., Orlando, R., and Sharp, J. S. (2017) A multivariate mixture model to estimate the accuracy of glycosaminoglycan identifications made by tandem mass spectrometry (MS/MS) and database search. *Mol. Cell. Proteomics* **16,** 255–264
17. Deutsch, E. (2008) mzML: a single, unifying data format for mass spectrometer output. *Proteomics* **8,** 2776–2777
18. Chambers, M. C., *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920
19. Bald, T., Barth, J., Niehues, A., Specht, M., Hippler, M., and Fufezan, C. (2012) pymzML–Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics* **28,** 1052–1053
20. Kaur, P., and O'Connor, P. B. (2006) Algorithms for Automatic Interpretation of High Resolution Mass Spectra. *J. Am. Soc. Mass Spectrom.* **17,** 459–468
21. Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., and Gambin, A. (2013) BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Anal. Chem.* **85,** 1991–1994
22. Prabhu, A., Venot, A., and Boons, G. J. (2003) New set of orthogonal protecting groups for the modular synthesis of heparan sulfate fragments. *Organic Letters* **5,** 4975–4978
23. Huang, Y., Mao, Y., Zong, C., Lin, C., Boons, G., and Zaia, J. (2015) Discovery of a heparan sulfate 3-O-sulfation specific peeling reaction. *Anal. Chem.* **81,** 592–600
24. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.* **44(D1),** D447–D456
25. Maxwell, E., Tan, Y., Tan, Y., Hu, H., Benson, G. Aizikov, K., Conley, S., Staples, G. O., Slysz, G. W., Smith, R. D., and Zaia, J. (2012) GlycReSoft: A software package for automated recognition of glycans from LC/MS data. *PLoS ONE* **7,** e45474