A Multivariate Mixture Model to Estimate the Accuracy of Glycosaminoglycan Identifications Made by Tandem Mass Spectrometry (MS/MS) and Database Search

Yulun Chiu‡§, Paul Schliekelman¶, Ron Orlando‡, and Joshua S. Sharp||**

We present a statistical model to estimate the accuracy of derivatized heparin and heparan sulfate (HS) glycosaminoglycan (GAG) assignments to tandem mass (MS/MS) spectra made by the first published database search application, GAG-ID. Employing a multivariate expectation-maximization algorithm, this statistical model distinguishes correct from ambiguous and incorrect database search results when computing the probability that heparin/HS GAG assignments to spectra are correct based upon database search scores. Using GAG-ID search results for spectra generated from a defined mixture of 21 synthesized tetrasaccharide sequences as well as seven spectra of longer defined oligosaccharides, we demonstrate that the computed probabilities are accurate and have high power to discriminate between correctly, ambiguously, and incorrectly assigned heparin/HS GAGs. This analysis makes it possible to filter large MS/MS database search results with predictable false identification error rates.

Molecular & Cellular Proteomics 16: 10.1074/mcp.M116.062588, 255–264, 2017.

Heparin and heparan sulfate (HS), members of the glycosaminoglycan (GAG) family, are linear polysaccharides composed of repeating disaccharide building blocks of variously sulfated hexuronic acid (1→4) 6-glucosamine units that structurally differ solely by the length of the oligosaccharide and degree of modification, with heparin being more heavily sulfated and having less N-acetylation than HS. Interacting with proteins, heparin/HS play essential roles in a wide variety of biological processes, including anticoagulation (1), cell proliferation (2, 3), and carcinogenesis (4, 5). The specificity of these interactions is driven by the pattern of modification of heparin/HS oligosaccharide sequences. To understand the molecular role of heparin/HS, it is necessary to correlate function with the fine structure of the carbohydrate. However, the non-template-driven biosynthesis of heparin/HS results in extremely diverse structures. Analyzing heparin/HS is challenging for three reasons: the presence of multiple isomeric sequences in a complex mixture of oligosaccharides, the difficulty of separating the isomers, and the facile loss of sulfates in MS/MS (6).

We previously introduced a method for structurally sequencing heparin/HS oligosaccharides that involves chemical derivatizations to replace labile sulfates with stable acetyl groups (7). This derivatization scheme allows for the use of reverse-phase liquid chromatography (LC) for high-resolution separation of isomeric heparin/HS oligosaccharides and MS/MS for sequencing them. However, the data from this derivatization method cannot be easily incorporated into current glycomic software, such as GlycoWorkbench (8), due to the multistep derivatizations and lack of a scoring algorithm that accurately evaluates the matches. We recently reported the development of a software tool for the high-throughput analysis of LC-MS/MS data from these derivatized heparin/HS oligosaccharides, entitled GAG-ID (9), which is the first database-driven software package for this purpose. GAG-ID produces a GAG sequence assignment for each input spectrum; however, some assignments are true matches and some are false. False matches arise from low-quality spectra, a simplified scoring scheme, and the presence of isomeric sequences with very similar theoretical fragmentation spectra coupled with incomplete or ambiguous MS/MS spectra (10). Thus, a statistical approach for evaluating and reporting the confidence associated with the GAG sequence assignment is required.

Heparin/HS oligosaccharides have linear structures composed of a finite, defined array of disaccharide sequences, analogous to peptide sequences. We have modeled our strategy on approaches currently used to analyze peptide MS/MS data. Tandem mass spectrometry has become the method of choice for high-throughput protein identification and quantification in most large-scale studies (11, 12). The most common approach to peptide identification from MS/MS spectra is database searching (13). In this approach, experimental MS/MS spectra are annotated by theoretically derived spec-
Molecular & Cellular Proteomics 16.2

A Multivariate Mixture Model for GAG Validation

tra predicted by peptides contained in a protein sequence database. Several database search tools are available, including SEQUEST (14), MASCOT (15), X!TANDEM (16), and others (17–20). A current challenge for high-throughput proteomics is to use database search results from large numbers of MS/MS spectra to derive a list of identified peptides and their corresponding proteins. However, not all peptide assignments are correct. Incorrect identifications arise from many sources, including the peptide sequence not being in the search database or the MS/MS spectra quality being too poor to be used for the database search (21). In addition, this task necessarily entails distinguishing correct peptide assignments from false identifications among database search results. In the case of small datasets, this can be achieved by researchers with expertise in manually verifying the peptide assignment to spectra made by a database search program. However, this time-consuming approach is not feasible when such expertise is not available or when analyzing large high-throughput datasets that contain tens of thousands of spectra.

Alternatively, researchers can attempt to separate the correct from incorrect peptide assignments by applying filtering criteria on the basis of database search scores. However, the number of false positives and false negatives that result from applying such filters is not known nor is it known how those findings are affected by the mass spectrometer, sample preparation, or spectrum quality. Furthermore, researchers often use different filtering criteria, making it particularly difficult to compare the results obtained by different research groups.

Several approaches have been introduced to translate the output scores from peptide database searches into probabilities or estimated false discovery rates. These can be roughly divided into single-spectrum (19, 22) and global (whole datasets) modeling approaches (23, 24). The most commonly used single-spectrum statistical measure is the expectation value, a calculation of the expected number of peptides with scores equal to or better than the observed score under the assumption that the peptide is assigned to the experimental spectrum by random chance. However, the conversion of a raw search score into an expectation value does not control the overall identification error rates, as its construction does not specifically involve steps to correct for multiple testing. In contrast, global modeling approaches are concerned with constructing the distribution of the search scores by taking the top-scoring peptide assignment for each experimental spectrum in the whole dataset. The global modeling approaches can be broadly grouped into two categories: the target-decoy approach (TDA) and the probability-based approach.

TDA—This group of methods relies on searching target and decoy databases and computing an optimized cutoff score for each dataset to reliably differentiate between correct and incorrect results (25). This approach involves two steps. In the first step, MS/MS spectra are searched against a target dataset of protein sequences augmented with the reverse (or randomized or shuffled) sequences of the same database. The approach assumes that the same distribution is followed for matches to the decoy peptide sequences and false matches to sequences from the target database. In the second step, peptide assignments are filtered using various score cutoffs, and the corresponding false discovery rate for each cutoff is estimated. The advantage of this false discovery rate estimation method is that it is simple to implement and requires minimal distributional assumptions, which makes it easily applicable in a variety of situations. A drawback to this approach includes doubling the search time. In addition, a critical issue arises as to whether reversing or randomizing sequences can provide an accurate assessment of the distribution of false peptide matches when many of those sequences are known to be homologous to the true peptides rather than completely random (26).

Probability-Based Approach—The probability-based method is exemplified by PeptideProphet (23), which models the distributions of database search scores and auxiliary information observed from all peptide assignments in the dataset as a two-component mixture of distributions representing correct and incorrect identifications. A discriminant search score, \( F \), is computed by combining multiple parameters related to the search score, including the search score, itself, \( Xcorr \), and its derivative, \( \Delta Cn \) score, in the case of SEQUEST search results. PeptideProphet generates the probability that the peptide assignment with discriminant score \( F \) is correct. These discriminant score distributions are modeled for each dataset using the expectation-maximization (EM)\(^1\) algorithm, leading to the posterior probability of correct identifications as inferential indicators. These probabilities are then used to estimate the false discovery rate for any minimum probability used as a cutoff. The limitations of PeptideProphet are largely related to the parametric assumptions and fixed coefficients when computing the discriminant search score.

The development of software tools in glycomic research is undergoing rapid changes, yet the tools available for high-throughput glycomics lag behind those developed for their protein counterparts, especially in the field of GAGs (27). Two software tools have recently been described for sequencing heparin/HS oligosaccharides by MS/MS. Hu et al. (28) published HS-SEQ, the first comprehensive algorithm for de novo HS sequencing using high-resolution negative electron transfer dissociation tandem mass spectra. That program aims to optimize the sulfation patterns of GAG oligosaccharides without searching against databases, which do not exist because of the nontemplate nature of GAG biosynthesis. Hu et al. described impressive success with their de novo approach, but HS-SEQ lacks a statistical model for assessing the confidence associated with the reported GAG identifications, re-

\(^1\) The abbreviations used are: EM, expectation-maximization; GAG, glycosaminoglycan; HS, heparan sulfate; LC, liquid chromatography; MS/MS, tandem mass spectrometry; S-\( \Delta \)Dev (%), delta deviation; TDA, target–decoy approach.
lying upon expert curation of spectra and validation of results. We reported a database-driven search approach, GAG-ID, which, to our knowledge, is the first software package capable of analyzing LC-MS/MS data of derivatized HS oligosaccharides. A database-driven approach to HS sequencing allows for the quick analysis of large datasets of LC-MS/MS runs of smaller oligosaccharides (tetrosaccharides to dodecasaccharides), which should facilitate the statistical modeling of the results for automated validation. In this work, we test global statistical approaches to modeling GAG-ID data and describe a robust and accurate statistical model to assess the validity of GAG identifications made by MS/MS and database searching. Employing database search scores and measures of the preference of the most likely identification to the second most likely identification (see Multivariate EM Model under “Experimental Procedures”), the method applies machine learning techniques to distinguish correct from ambiguous and incorrect assignments in the dataset. In doing so, the method computes a posterior probability of a correct match for each GAG spectrum assignment. We apply this method to GAG-ID database search results for LC-MS/MS spectra with the GAG-ID search program (Fig. 1) using GAG-DB, which contains all possible theoretical HS sequences, including 992 tetramers, 15,872 hexamers, 253,952 octamers, 4,083,232 decamers, and 65,011,712 dodecamers. To execute the TDA, therefore, one needs a mechanism to create decoy sequences that are not biologically possible, with each method producing a decoy database of the same size (i.e., same number of oligosaccharide sequences) and same composition (i.e., same quantity of each building block for each sequence). However, because our target database contains every possible combination of heparin/HS disaccharides, we cannot merely reverse, shuffle, or randomize the disaccharides in our target sequence database to form our decoy database. Therefore, we investigated several alternative ways to form a decoy database. Fig. 2a shows an example of a tetrasaccharide target sequence. We introduced several approaches to generate an HS decoy database from a target database. (1) Decoy misplaced method 1 (BABA): We flipped each disaccharide unit (Fig. 2b), resulting in each oligosaccharide beginning with a glucosamine unit as opposed to the uronic acid unit found in the target database. (2) Decoy misplaced method 2 (AABB): We shuffled the order of the oligosaccharide (Fig. 2c), with two uronic acid units followed by two glucosamine units, as opposed to the uronic acid–glucosamine repeat found in the target database. (3) Decoy m/z-shift method 1 (AABB47): We added an artificial mass of 47 Da to each uronic acid unit and subtracted from each glucosamine unit on the reducing end to keep the same number of spectra that passed the MS1 tolerance filter with decoy database searching (Fig. 2d). (4) Decoy m/z-shift method 2 (ABAB47): We added an artificial mass of 47 Da to the uronic acid on the nonreducing end and subtracted from the glucosamine unit on the reducing end (Fig. 2e). We applied each of these methods to all possible GAG tetrasaccharide sequences, resulting in decoy databases of the exact same size as the target database and intact masses identical to sequences in the target database, with sequences that are biologically impossible; therefore, all hits to any of these decoy databases represent incorrect matches.

**Experimental Procedur**

**Data**—Tandem mass spectra used in this study were generated from LC-MS/MS runs on a defined mixture of 21 tetrasccharides as well as seven longer defined oligosaccharides, as previously described (9). For each LC-MS/MS run, a defined sample processed with a serial chemical derivatization was analyzed by C18 reverse-phase LC-MS/MS using an LTQ-FT (Thermo-Finnigan, San Jose, CA). Oligosaccharide assignments with known validity were obtained by searching these spectra with the GAG-ID search program (Fig. 1) using GAG-DB, which contains all possible theoretical HS sequences, including 992 tetramers, 15,872 hexamers, 253,952 octamers, 4,083,232 decamers, and 65,011,712 dodecamers. To execute the TDA, therefore, one needs a mechanism to create decoy sequences that are not biologically possible, with each method producing a decoy database of the same size (i.e., same number of oligosaccharide sequences) and same composition (i.e., same quantity of each building block for each sequence). However, because our target database contains every possible combination of heparin/HS disaccharides, we cannot merely reverse, shuffle, or randomize the disaccharides in our target sequence database to form our decoy database. Therefore, we investigated several alternative ways to form a decoy database. Fig. 2a shows an example of a tetrasaccharide target sequence. We introduced several approaches to generate an HS decoy database from a target database. (1) Decoy misplaced method 1 (BABA): We flipped each disaccharide unit (Fig. 2b), resulting in each oligosaccharide beginning with a glucosamine unit as opposed to the uronic acid unit found in the target database. (2) Decoy misplaced method 2 (AABB): We shuffled the order of the oligosaccharide (Fig. 2c), with two uronic acid units followed by two glucosamine units, as opposed to the uronic acid–glucosamine repeat found in the target database. (3) Decoy m/z-shift method 1 (AABB47): We added an artificial mass of 47 Da to each uronic acid unit and subtracted from each glucosamine unit on the reducing end to keep the same number of spectra that passed the MS1 tolerance filter with decoy database searching (Fig. 2d). (4) Decoy m/z-shift method 2 (ABAB47): We added an artificial mass of 47 Da to the uronic acid on the nonreducing end and subtracted from the glucosamine unit on the reducing end (Fig. 2e). We applied each of these methods to all possible GAG tetrasaccharide sequences, resulting in decoy databases of the exact same size as the target database and intact masses identical to sequences in the target database, with sequences that are biologically impossible; therefore, all hits to any of these decoy databases represent incorrect matches.

**Multivariate Mixture Model**—We divided the assigned spectra into three categories: correct assignments, incorrect assignments, and ambiguous assignments. Correct assignments are defined as those with a high GAG-ID score and for which the second-best assignment has a substantially lower score. Incorrect assignments are defined as assignments with a low GAG-ID score. The “ambiguous” assignments are defined as those with a high GAG-ID score but for which the second-best match has a score that is nearly as high. The nature of glycomics and the limitation of analytic separation require us to consider the second-best match in order to differentiate isomeric structures when they are present. This is quantified using the S–ΔDev (%) values. The S–ΔDev (%) is defined below, where S″TOP 1 and S″TOP 1 are the highest and second highest scores of the GAG sequence used to annotate the spectrum.

\[
S - \Delta\text{Dev}(\%) = 100 \times \frac{S_{\text{TOP} 1} - S_{\text{TOP} 2}}{S_{\text{TOP} 1}}
\] (Eq. 1)

The larger the S–ΔDev (%), the more significant the best assignment is compared with the second-best assignment.
We developed a likelihood model in which we assume that the GAG-ID scores (G) and the Delta deviation (D) for a spectrum follow a joint normal distribution. The mean and variance depend on the category in which the assignment falls (correct, incorrect, or ambiguous). The overall distribution is thus

We developed a likelihood model in which we assume that the GAG-ID scores (G) and the Delta deviation (D) for a spectrum follow a joint normal distribution. The mean and variance depend on the category in which the assignment falls (correct, incorrect, or ambiguous). The overall distribution is thus
\[ p(G,D) = p(G,D|+) p(+) + p(G,D|?) p(?) + p(G,D|--) p(-) \]  

(Eq. 2)

The probabilities \( p(G,D|+) \), \( p(G,D|?) \), and \( p(G,D|--) \) are the probabilities of a spectrum having the GAG-ID score of G and the S-∆Dev value of D when the assignment is correct, ambiguous, and incorrect, respectively. The quantities \( p(+) \), \( p(?) \), and \( p(-) \) are the respective prior probabilities for the assignment being correct, ambiguous, or incorrect. Equation 2 is known as a mixture distribution. The probability for the pair \((G,D)\) depends on two random factors: 1) whether the assignment is correct, ambiguous, or incorrect and 2) the probability of the observed values \((G,D)\) given the assignment category. After fitting this model, we can calculate the posterior probabilities for the assignment categories

\[ p(+|G,D) = \frac{p(G,D|+) p(+) + p(G,D|?) p(?) + p(G,D|--) p(-)}{p(G,D)} \]  

(Eq. 3)

We use the multivariate EM algorithm (29) to estimate model parameters. The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given dataset when the dataset has missing values or unobserved variables. In this case, the assignment category \((+,-, or ?)\) of each \((G,D)\) pair is unobserved, making regular likelihood maximization intractable. The EM algorithm allows us to maximize the likelihood and fit the parameters in Equation 2 even in the absence of this information.

The EM algorithm involves two steps, the expectation (E) step and the maximization (M) step. These steps are repeated iteratively. The application of the EM to the case where the components follow a normal distribution is well known and can be easily adapted to our setting. When the components of the mixture follow a normal distribution, the E and M steps are combined into one set of equations. The estimate in the \((j+1)\)th iteration for the prior probability that an assignment is correct is given by

\[ p_{j+1}(+) = \frac{1}{\sum_{i=1}^{N} p_{j}(+|G_{i}, D_{i})} \]  

(Eq. 4)

where \(G_{i}\) is the GAG-ID score and \(D_{i}\) is the Delta deviation for the sequence assignment to spectrum \(i\) and \(p_{j}(+|G_{i}, D_{i})\) is the probability, calculated from the parameters estimated in the \(j\)th iteration that the assignment is correct given \((G_{i}, D_{i})\). We can calculate \(p_{j}(+|G_{i}, D_{i})\) as

\[ p_{j}(+|G_{i}, D_{i}) = \frac{p_{j}(G_{i}, D_{i}|+) p(+) + p_{j}(G_{i}, D_{i}|?) p(?) + p_{j}(G_{i}, D_{i}|--) p(-)}{p_{j}(G_{i}, D_{i})} \]  

(Eq. 5)

The GAG-ID scores and S – ∆Dev (\%) values are modeled as bivariate normal distributions, conditional on the assignment status

\[ p_{j}(G_{i}, D_{i}|+) = \frac{1}{(\sigma_{G}(+)^{2} + \sigma_{D}(+)^{2})} \exp\left(-\frac{(G_{i} - \mu_{G}(+))^{2}}{2 \sigma_{G}(+)^{2}}\right) \]  

(Eq. 6)

where \(\mu_{G}(+, +)\) and \(\sigma_{G}(+, +)\) are the estimates in the \(j\)th iteration for the mean and standard deviation in the GAG-ID among correctly assigned spectra. Likewise, \(\mu_{D}(+, +)\) and \(\sigma_{D}(+, +)\) are the estimates in the \(j\)th iteration for the mean and standard deviation in S – ∆Dev among correctly assigned spectra. In Equation 5, we assume independence between G and D. Similar expressions are defined for the ambiguous and negative assignments.

The estimates in the \(j\)th iteration for \(\mu_{G}\) and \(\sigma_{G}\) are given by

\[ \mu_{G}(j + 1, +) = \frac{1}{\sum_{i=1}^{N} p_{j}(+|G_{i}, D_{i})} \sum_{i=1}^{N} p_{j}(+|G_{i}, D_{i}) G_{i} \]  

(Eq. 7)

\[ \sigma_{G}(j + 1, +) = \left(\frac{1}{\sum_{i=1}^{N} p_{j}(+|G_{i}, D_{i})} \sum_{i=1}^{N} p_{j}(+|G_{i}, D_{i}) (G_{i} - \mu_{G}(j, +))^{2}\right)^{1/2} \]  

(Eq. 8)

The expressions for \(\mu_{D}(j + 1, +)\) and \(\sigma_{D}(j + 1, +)\) are identical but with \(\mu_{G}(j, +)\) and \(\sigma_{G}(j, +)\) replaced by \(\mu_{D}(j, +)\) and \(\sigma_{D}(j, +)\). Similar expressions apply for ambiguous and incorrect assignments. Starting from randomly chosen initial parameter values, Equations 4, 6, and 7 are applied iteratively until the change in parameters between iterations is less than a predefined value. Under suitable conditions, this process will converge to the maximum likelihood estimates (29).

On termination of the algorithm, final probabilities that sequences assigned to spectra are correct are computed according to Equation 3. By default, spectra are assigned to the category \((+, ?, -)\) that has the highest probability. However, the output also includes the probability that the assignment is correct \(p_{\text{correct}}\) or that the assignment is ambiguous \(p_{\text{ambiguous}}\). Thus, users can determine the appropriate cutoff.

One issue with the use of maximum likelihood for parameter estimation is the risk that the algorithm returns a local maximum in the likelihood surface rather than the global maximum, thus yielding false parameter estimates. In order to guard against this, we run the EM algorithm 1000 times with random starting values. If there are multiple peaks in the likelihood surface, then different starting values will tend to lead to optimization at the different peaks. We take the one that produces the highest likelihood as the global optimum. We did observe three different solutions among our 1000 runs. However, one of the three peaks had a likelihood value of approximate 1500 times the second highest value, and thus, we are confident that this is the correct global maximum.

The implementation software is written in R programming language and utilizes the mixtools package (version 1.0.2),
which is publicly available from Comprehensive R Archive Network (CRAN) (http://cran.r-project.org/). Density plots were generated using the density function in R. This function generates kernel density estimates of the probability distribution underlying a dataset.

Note that S-ΔDev values are percentages and therefore do not follow a normal distribution in reality. However, our data show that S-ΔDev values are crucial in correctly classifying assignments, and our goal in including S-ΔDev values in the likelihood is to provide an automated way to do this. Our results show that this approach performs adequately.

RESULTS

We have used a database-driven approach, GAG-ID, to successfully sequence a defined mixture of 21 synthesized tetrasaccharides. This approach scores ion matches of oligosaccharide fragments against theoretical fragmentation patterns generated from derivatized HS GAG sequences. A total of 18,682 MS/MS spectra were acquired, of which 1306 MS/MS spectra were identified as putative HS oligosaccharide sequences within specified MS tolerance. Of these MS precursors, 1295 MS/MS spectra were observed to have a GAG-ID score higher than 0, indicating that each spectrum in these 1295 MS/MS spectra is annotated with at least a structure. From these 1295 annotated MS/MS spectra, GAG-ID produced a GAG assignment for each input spectrum, some of which might be true matches and some false. For this purpose, the two most commonly used statistical validations for high-throughput proteomics data analysis, TDA and the EM algorithm, were used to examine the GAG-ID sequencing results.

TDA Underestimated the Error Rate in GAG-ID—TDA is well established in peptide sequencing in the field of proteomics. However, it is untested in the field of glycomics, GAG sequencing in particular. For peptide sequencing, this approach is simple to understand, easy to implement, and can be generally applied to all datasets and search engines. However, it is not clear how this approach can be applied to GAG data (see Decoy GAG Sequence in the Discussion). We generated decoy sequences from a putative HS GAG sequence by five different methods. The 18,682 MS/MS spectra generated from a defined mixture of 21 synthesized tetrasaccharide sequences were searched by GAG-ID against a target database in which the correct 21 synthesized tetrasaccharides were removed and then against a decoy database in which 21 analogous decoy tetrasaccharides were removed. Because the correct annotations were missing from the target database, all the assignments were false positives. If the decoy database is correctly constructed, it should produce a number of false positive matches that is similar to that of the target sequence. Fig. 3 illustrates that the number of mismatched spectra for all of the decoy sequences was substantially lower than that for the target sequence. Thus, all these decoy methods systematically underestimate the error rate.

TDA Underestimated the Error Rate by TDA. The number of incorrect identifications from GAG-ID plotted by searching against the target sequence (ABAB) database that is missing all correct target sequences (dotted blue line, the upper most curve) and decoy sequence (BABA, ABAB47, A4BAB4, AABB) databases of identical size to the edited target sequence database (the remaining lines). “A” represents uronic acid; “B” represents glucosamine. ABAB47 indicates that artificial mass of 47 Da was added to each uronic acid unit and subtracted from each glucosamine unit. A4BAB4 indicates that artificial mass of 4 Da was added to the uronic acid on the nonreducing end and subtracted from the glucosamine unit.

Three Normal Distributions of Correct, Ambiguous, and Incorrect HS Sequence Assignment—Next, we applied our likelihood-based approach to classify assignments as correct or not. The spectra were searched using GAG-ID against the complete theoretical putative tetrasaccharide database. In order to make a realistic example, no data cleaning (e.g. removal of poor quality spectrum) was applied. Because we knew the identities of the 21 sequences, we could classify all the assignments with certainty.

Initially, we used a two-class assignment of correct and incorrect. However, the resulting fitted distribution was not consistent with the observed data (see Supplemental Fig. S1). There appeared to be three peaks to the distribution. Fig. 4 shows density plots of the GAG-ID scores for the whole dataset (solid blue line), correct assignments (solid green line), incorrect assignments (solid black line), and ambiguous assignments (solid red line).

From the results of the EM model for the three normal distributions and prior knowledge of which assignments were correct, we determined that the peak with the highest GAG-ID score consisted of correct assignments (Fig. 4, green dotted line) and the peak with the lowest GAG-ID score consisted of incorrect assignments (Fig. 4, black dotted line); however, the middle peak, which is defined as the GAG-ID score from 2 to 4, reflected an “ambiguous” category between correct and incorrect assignments. Manual examination of the middle of the three peaks revealed that itcontains a total of 257 MS/MS spectra that are dominated by the correct category and ambiguous category: 118 of 257 MS/MS spectra are correctly assigned, including 73 singly charged spectra and 45 doubly charged spectra; 91 of 257 MS/MS spectra are ambiguously assigned, including seven singly charged spectra and 84 doubly charged spectra; and 48 of 257 MS/MS spectra are incorrectly assigned.

Multivariate Normal Distribution—To resolve the significant overlap of the three observed distributions in the single-vari-
correct and incorrect assignments of very low Delta deviation. Correct group, confidently incorrect group, and a mixture of incorrect assignments, dividing the data into a confidently Delta deviation grouping among correct assignments and the false negative rate is 0.04. For the ambiguous distribution, 5 of 130 spectra are assigned correctly, and the false negative rate is 0.04. For the ambiguous distribution, 339 of 503 spectra are assigned correctly, and 461 of 503 are assigned correctly within the top two hits. These results show that our approach is highly effective at categorizing assignments and that ambiguous assignments result primarily from cases where there are multiple database hits with similar scores.

**DISCUSSION**

A primary bottleneck for dissemination of LC-MS/MS glycomic sequencing methods is the complexity of the data. There is currently no gold standard of features to be used to score glycomic data and examine the validity of search results. Here, we tested the utility and validity of using a multivariate EM model to measure the reliability
of search results from our GAG-ID algorithm using a comprehensive target database that contains all possible sequences in conjunction with a defined mixture of heparin/HS tetrasaccharides.

Decoy GAG Sequence—In the proteomic field, peptide sequence reversal is by far the simplest and most widely used method for creating decoy sequences. Peptide sequence reversal has two main advantages. First, because it preserves the general features of the target sequence list, reversed peptide sequences share the same degree of interprotein redundancy as the input target sequences. Second, as peptide sequence reversal represents a defined transformation, multiple research groups can generate the same decoy sequences. However, a GAG sequence consists of a repeating disaccharide unit structure. Due to the palindromic or symmetric fragmentation pattern, which does not strictly represent a null random distribution, the method of simply reversing the monosaccharide sequence is not suitable for creating a decoy counterpart (Fig. 3). Due to the comprehensive nature of the target database, a reversal (or randomiza-
tion) of the order of disaccharide repeats is already represented in the target database and is therefore unavailable for the decoy database. In silico modification of the disaccharide repeat units to render the non-biological sequences results in theoretical fragmentation spectra that are too far removed to serve as useful decoys for TDA analyses.

Ambiguous Assignment—The algorithm places spectra into one of three categories: The correct category is spectra with a high GAG-ID score (indicating that the highest scoring match in the database was a good match to the spectrum) and high S-ΔDev score (indicating that there were no other matches with similarly high score). The incorrect category is spectra with low GAG-ID scores, indicating cases where there was no tetrasaccharide in the database that was a good match. S-ΔDev scores for incorrect matches vary from moderate to high. This is because scores for incorrect matches are all drawn from the same null distribution and the second-best incorrect match does not have a tendency to either be similar or dissimilar to the best incorrect match. The ambiguous category is characterized by spectra with a low S-ΔDev and a range of GAD-ID scores. This category is dominated by doubly charged spectra where one of the top two matches is the correct one (see Figs. 7 and S4).

Nesvizhskii et al. (10) enumerated reasons for the occurrence of false identifications in database search results generated by tandem mass spectrometry, including a simplified scoring scheme, low quality spectra, fragmentation of multiple peptide ions, a restricted database search, incorrectly determined charge state or peptide mass, sequence variants, and the analysis of new peptides. Furthermore, the heterogeneity of the GAGs not only increases the difficulty of sample separation but also leads to problems in the analysis. As all possible sequences are present in the target database, any ambiguity in the MS/MS spectra (e.g. missed glycosidic bond cleavage, artificial signal that shares an m/z with a theoretical fragment ion) can result in an ambiguous assignment. Additionally, the presence of isomeric structures in a complex mixture of oligosaccharides can generate chimeric MS/MS spectra (MS/MS spectra of two or more isobaric molecules fragmenting simultaneously), resulting in an incorrect assignment for the best hit from the sequencing results, as well as assignments that, while possibly correct, have poor Delta deviations separating the top assignment from the second-best result. Therefore, the data are best described using a third distribution that is located between the prototypically correct and incorrect distributions that represent the spectra of heparin/HS oligosaccharides. In this third distribution, the top hit may or may not be correct but cannot be reliably differentiated from one or more other top hits due to the high level of degeneracy that is an unavoidable characteristic of the comprehensive database approach. We labeled this third

### Table I

| Spectrum # | GAG-ID Results | Observed Precursor m/z | GAG-ID Score | S-ΔDev(%) | Probability | Probability |
|------------|----------------|------------------------|--------------|-----------|-------------|-------------|
| 1          | HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1408.00 | 10.834 | 100 | 0.999999 | 0.000001 |
| 2          | HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1175.00 | 2.376 | 100 | 0.999873 | 0.000127 |
| 3          | HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1408.00 | 11.7845 | 0 | 0.002799 | 0.997200 |
| 4          | d-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1408.00 | 12.6687 | 0 | 0.001772 | 0.998227 |
| 5          | HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1175.00 | 5.985 | 0 | 0.019164 | 0.980835 |
| 6          | HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-GlcNS-HexA-AnMan | 1175.00 | 5.305 | 0 | 0.021271 | 0.978728 |

The correct assignment for spectra 3 through 5 is the sequence of spectrum 1; the correct assignment for spectra 6 and 7 is the sequence of spectrum 2.
distribution the ambiguous distribution. These results lead us to conclude that a proper model includes three component distributions, which we illustrate in Fig. 6a: incorrect (red dots), ambiguous (green dots), and correct (blue dots).

Our study is based on data collected from a mixture of 21 synthesized tetrasaccharide sequences. We must address some practical issues when applying this approach to high-throughput heparanomic analyses, in which we may encounter large GAGs with complex structures. One issue is how to set up the Delta deviation (S−ΔDev (%)) cutoff. The Delta deviation is the distance between the best assignment and the second-best assignment. A higher Delta deviation is suggested to differentiate the isomers. This study employed a multivariate mixture model to estimate the accuracy of the GAG-ID assignment results. We plan to make the univariate (GAG-ID score) EM model with two normal distributions available as an option for a future study. The ultimate goal to annotate real samples may require concomitant efforts from the experimental part (extraction and separation) and the computational part (search algorithm and accuracy estimation). In this paper, we addressed the EM model regarding estimation accuracy. In fact, this is computationally more difficult because it is still a challenging problem to determine the number of GAGs and related isomers that exist. For example, no method has been proposed thus far to estimate the accuracy of correct assignment of GAG identification. We plan to study these computational problems in the future.

Acknowledgments—We thank Prof. David L. Tabb for helpful discussions and critical insights, and Prof. William York for critical remarks. We also thank LeeAnn Chastain (MDACC) for editing assistance.

REFERENCES

1. Iozzo, R. V., and San Antonio, J. D. (2001) Heparan sulfate proteoglycans: Heavy hitters in the angiogenesis arena. J. Clin. Invest. 108, 349–355
2. Holt, C. E., and Dickson, B. J. (2005) Sugar codes for axons? Neuron 46, 160–172
3. Dityatev, A., and Schachner, M. (2003) Extracellular matrix molecules and synaptic plasticity. Nat. Rev. Neurosci. 4, 456–468
4. Muramatsu, T., and Muramatsu, H. (2008) Glycosaminoglycan-binding cytokines as tumor markers. Proteomics 8, 3350–3359
5. Knelson, E. H., Nee, J. C., and Blobe, G. C. (2014) Heparan sulfate signaling in cancer. Trends Biochem. Sci. 39, 277–288
6. Peng, Y., and Lindahl, U. (2013) Molecular diversity of heparan sulfate. J. Clin. Invest. 128, 169–173
7. Huang, R., Liu, J., and Sharp, J. S. (2013) An approach for separation and complete structural sequencing of heparin/heparan sulfate-like oligosaccharides. Anal. Chem. 85, 5787–5795
8. Ceroni, A., Maas, K., Geyer, H., Geyer, R., Dell, A., and Haslam, S. M. (2008) GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans. J. Proteome Res. 7, 1650–1659
9. 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
10. Nesvizhskii, A. I., Vitik, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods 4, 787–797
11. Hijrhayashi, J., Hashidate, T., Arata, Y., Nishi, N., Nakamura, T., Hiroshima, M., Urashima, T., Oka, T., Futai, M., Muller, W. E., Yagi, K., and Kasai, K. (2002) Oligosaccharide specificity of galectins: A search by frontal affinity chromatography. Biochim. Biophys. Acta 1572, 232–254
12. Wuthrer, M., Deelder, A. M., and Hokke, C. H. (2005) Protein glycosylation analysis by liquid chromatography-mass spectrometry. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 825, 124–133
13. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discov. Today 9, 173–181
14. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989
15. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probabilistic-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567
16. Craig, R., and Beavis, R. C. (2004) TANDEM: Matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467
17. Fridman, T., Razumovskaya, J., Verberkmoes, N., Hurst, G., Protopopescu, V., and Xu, Y. (2005) The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. J. Bioinform. Comput. Biol. 3, 455–476
18. Sadygov, R. G., Liu, H., and Yates, J. R. (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal. Chem. 76, 1664–1671
19. Geer, L. Y., Markey, S. P., Kowalek, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. J. Proteome Res. 3, 958–964
20. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProIDb: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2, 1406–1422
21. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658
22. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identification using general scoring schemes. Anal. Chem. 75, 768–774
23. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392
24. Choi, H., and Nesvizhskii, A. I. (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J. Proteome Res. 7, 254–265
25. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214
26. Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J. Proteome Res. 7, 29–34
27. Kalliemia, M. J., Ruhaak, L. R., Lebrilla, C. B., and Amster, I. J. (2014) Oligosaccharide analysis by mass spectrometry: A review of recent developments. Anal. Chem. 86, 196–212
28. Hu, H., Huang, Y., Mao, Y., Xu, X., Yu, L., Liu, J., Zong, C., Boons, G.-J., Lin, C., Xia, Y., and Zia, J. (2014) A computational framework for heparan sulfate sequencing using high-resolution tandem mass spectra. Mol. Cell. Proteomics 13, 2490–2502
29. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 39, 1–38