research articles Journal of proteome research

Automatic Determination of *O*-Glycan Structure from Fragmentation Spectra

David Goldberg,* Marshall Bern, Bensheng Li, and Carlito B. Lebrilla

Palo Alto Research Center, Palo Alto, CA, 94301, and Department of Chemistry and School of Medicine, Biochemistry and Molecular Medicine, University of California, Davis, California 95616

Received February 3, 2006

Glycosylation is one of the most important classes of post-translational protein modifications, but the identification of glycans is difficult because of their branched structures and numerous isomers. We describe an algorithm called *CartoonistTwo* that proposes structures for O-linked glycans by automatically analyzing fragmentation mass spectra. *CartoonistTwo* improves upon previous glycan identification software primarily in its scoring function, which can more successfully distinguish among a number of similar structures. *CartoonistTwo* was designed and tested with FTICR mass spectra, and includes automatic recalibration and peak selection especially tuned for such data, yet it can be easily adapted to fragmentation spectra (MS^2 or MS^n) from other instrument types. On a validated test set of 34 SORI−CID MS^n FTICR spectra from *Xenopus* egg jelly, *CartoonistTwo* gave the manually determined structural assignment either the first or second highest score over 90% of the time. And for over 50% of these spectra, *CartoonistTwo* selected a *unique* highest scoring structure that agreed with the manually determined one.

Keywords: glycomics • O-linked oligosaccharides • MALDI−FTMS • cartoon • tandem mass spectrometry • MS/MS • FT−ICR • *O*-glycan.

1. Introduction

In recent years, identification of proteins by tandem mass spectrometry has become quite common, but the identification of post-translational modifications remains a difficult problem. Especially challenging is the identification of carbohydrate attachments (glycans), because carbohydrates are themselves polymeric molecules with the additional complexity that they are typically branched rather than linear structures. There is, however, ample reason to pursue improved identification of glycans by mass spectrometry. Glycosylation is perhaps the most important of all post-translational modifications. It has been estimated that over 50% of all eukaryotic proteins are glycosylated,[1] and it is well-known that the glycans presented on cell surfaces are vital for cell−cell communication.[2] Glycans define A/B/O blood groups and play roles in autoimmune disease and cancer, and the loss of a single glycosyltransferase has even been suggested as the crucial mutation that enabled human brain expansion.[3,4]

Manual interpretation of fragmentation mass spectra of complex molecules can be tedious and error prone. We have developed a tool called *CartoonistTwo* for rapid and automatic determination of the structures of detached O-glycans from fragmentation mass spectra. *CartoonistTwo* produces a ranked list of structures that best explain a spectrum, along with explanations for the ranking, and can be used to determine the most likely structure or to suggest which additional experiments (e.g., exoglycosidase experiments) are needed to identify the compound more definitively. The current version of *CartoonistTwo* determines structures only to the level of *cartoons*. A cartoon specifies the composition of monosaccharides and their connectivity (or *topology*), but not does not specify the type of bond (e.g., α or *β*) nor distinguish isomeric monosaccharides e.g., *N*-acetyl-D-galactosamine (GalNAc) and *N*-acetyl-D-glucosamine (GlcNAc).

CartoonistTwo generates all possible cartoons and ranks them by score, often with enough discrimination so that there is a unique highest-scoring cartoon. The high performance of *CartoonistTwo* stems from a detailed scoring algorithm with three well-founded innovations. First, the scorer uses low-intensity peaks, guarding against the inclusion of noise peaks by employing a statistical confidence measure based on both intensity and m/z. Second, the scorer uses not only the peaks present in the spectrum, but also peaks that are missing—fragments of a proposed structure that do not appear in the spectrum. Third, it assumes a model of low-energy glycan fragmentation, which we refer to as "shedding". In this model, monosaccharides are cleaved (or shed) from the glycan one at a time, with the charge remaining with the larger daughter ion. See Figure 1. This model successfully predicts the peaks actually observed in spectra produced by FTICR (Fourier Transform Ion Cyclotron Resonance) mass spectrometry employing either multiple rounds of SORI−CID (Sustained Off-Resonance Irradiation Collision-Induced Dissociation) or IRMPD (Infrared Multiphoton Dissociation) fragmentation.[5]

Several computer programs for identifying oligosaccharides by tandem MS—typically to the level of cartoons—have been described in the literature. GlycosidIQ,[6] GlycoMod,[7] and Gly-

* To whom correspondence should be addressed. Fax: (650) 812-4471. E-mail: goldberg@parc.com.

10.1021/pr060035j CCC: $33.50

Published on Web 05/17/2006

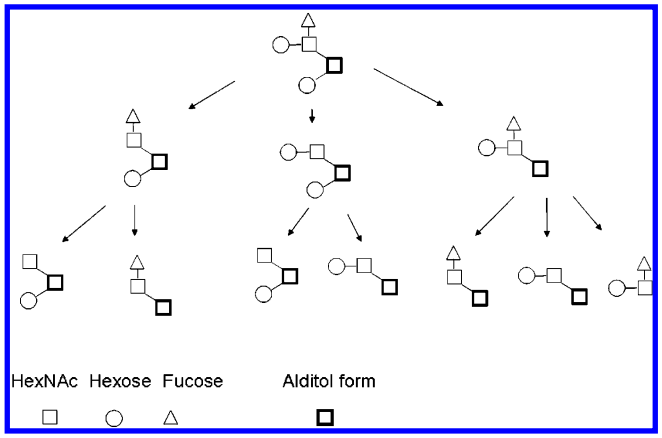

Figure 1. Cartoon at the top is a proposed parent structure, and underneath are some of the possible fragments produced by the shedding model. The fragments are arranged in a partial order, or *shedding tree,* with arrows connecting two structures that differ by a single monosaccharide.

coSearchMS[8] are database-search programs. Glycan databases[9−11] are much less complete than protein databases, especially in their coverage of *O*-glycans, and hence glycan database-search programs do not yet provide the same general utility as protein database-search programs like SEQUEST[12] and Mascot.[13] There are also several de novo identification programs. StrOligo[14,15] handles only N-glycans and exploits the special structure of those glycans (e.g., the trimannosyl core). STAT[16] and GLYCH[17] handle *O*-glycans as well as *N*-glycans, but have very basic scoring functions, and hence often return large numbers of equally good candidate structures, especially for spectra of larger glycans. GLYCH was initially demonstrated with only small, mostly linear, reference oligosaccharides, along with the higher-energy CID fragmentation used with MALDI−TOF experiments. Higher-energy CID produces some cross-ring fragmentation, and hence this program also goes beyond topology and attempts to identify linkage information.

2. Experimental Section

2.1. Methods. The experimental procedure has been described previously.[18] In brief, mucin-type O-linked oligosaccharides were detached and isolated from egg jelly glycoproteins of two frog species, *Xenopus laevis* and *Xenopus tropicalis*. The procedure converts the reducing ends of the oligosaccharides to alditols, which confers the side benefit of mass labeling the reducing-end monosaccharide with an increase of 2.0156 Da (two hydrogen atoms). All spectra were obtained on a commercial MALDI−FTICR instrument (Ion Spec, Irvine, CA). The Na^+ concentration was enriched to produce primarily singly charged, sodiated species. Two types of fragmentation were used: IRMPD and SORI−CID MS^n. Both of these "slow-heating" methods add energy more gradually than the CID methods used with MALDI−TOF instruments, and hence almost exclusively access low-energy fragmentation pathways.[5,19,20] IRMPD produces individual spectra containing essentially the same peaks as a sequence of MS^n SORI−CID spectra, but with the advantages of faster duty cycles and less ion loss due to scattering.[19] Although *CartoonistTwo* has been run on both individual IRMPD spectra and sequences of MS^n SORI−CID data, the test set contained only sequences of SORI−CID spectra, as structures for these spectra were better validated. The test set is available by request from the authors.

2.2. Software. *CartoonistTwo* is a Unix shell script that invokes a series of three programs written in C. The first program processes the spectra by picking peaks that are likely to represent glycans and then assigning confidence values to these peaks. The second program is an enhanced version of the original *Cartoonist* program,[21] designed for single-MS MALDI−TOF spectra, which assigns possible glycan compositions to the peaks (specifying the identities of monosaccharides but not their connectivity) and recalibrates m/z measurements based on these tentative assignments. The third program in the series does the actual identification: it generates all plausible candidate topologies and scores them.

2.3. Algorithms. 2.3.1. Finding Significant Peaks. The first program in the series relies on statistical modeling of noise peaks. At first glance, FTMS spectra appear much noisier than spectra from lesser instruments, as seen in Figure 2. Impulse noise in the (cyclotron) frequency domain produces a lush lawn of noise peaks in the m/z domain. Previous papers have employed heuristics to select peaks, for example, selecting all peaks of intensity at least 5 times the baseline noise level,[22] which can be set using a histogram of peak intensities.[23] Here, we give a more principled method, which uses a peak histogram to compute a p value, the chance that a given peak would arise from noise alone.

The instrument software delivers a spectrum with 5000−30 000 picked peaks (local maxima), only a minuscule fraction of which are indeed signal (peaks representing glycan fragments). Our preprocessing software computes a peak histogram as shown in Figure 4 and then plots about 15 points (x, y), where x is the center intensity of a histogram bin and y is the logarithm of the number of peaks within that bin. Figure 4 shows linear and quadratic fits to these points. The quadratic fit is better, which is not surprising, as we might expect the intensity distribution to fall off as a normal distribution, that is, as $\exp(-x^2)$. Once we have fit the log frequency to a quadratic, $-a_0 - a_1 x - a_2 x^2$, then we can compute the probability that a noise peak has intensity at least α by

$$p(\alpha) = \int_\alpha^\infty e^{-(a_0+a_1x+a_2x^2)}dx = \frac{1}{\sqrt{a_2}}\, e^{-a_0+a_1^2/4a_2}\mathrm{erfc}\left(\left(\alpha + \frac{a_1}{2a_2}\right)\sqrt{a_2}\right)$$

If there are N peaks in the spectrum, then we would expect to see at least one noise peak of intensity greater than α with probability of $1 - (1 - p(\alpha))^N$, and hence we can set the *significance* of a peak of intensity α to be $(1 - p(\alpha))^N$. To handle variation in noise with m/z, we divide the spectrum into overlapping segments containing $N/4$ peaks and perform the analysis on each segment separately. The significance for a given peak is set by the segment for which it is closest to the center.

2.3.2. Recalibration. FTMS offers the most accurate mass measurement of any currently available instrument technology, yet the measurements can be greatly improved by recalibration using either internal chemical calibrants[24] or correlation of different charge states of the same species.[25] Here, we describe precise recalibration based on tentative peak assignments. This recalibration method has been used already in the scoring phase of de novo peptide sequencing,[26] but with glycans, it can even be used in the preprocessing phase, because it is possible to make tentative peak assignments (at the level of monosaccharide composition) without reference to a candidate parent molecule. As reported previously, *Cartoonist*[21] uses this method for analysis of single-MS MALDI−TOF glycan spectra. Here, we show how to extend the technique to fragmentation spectra and FTMS.

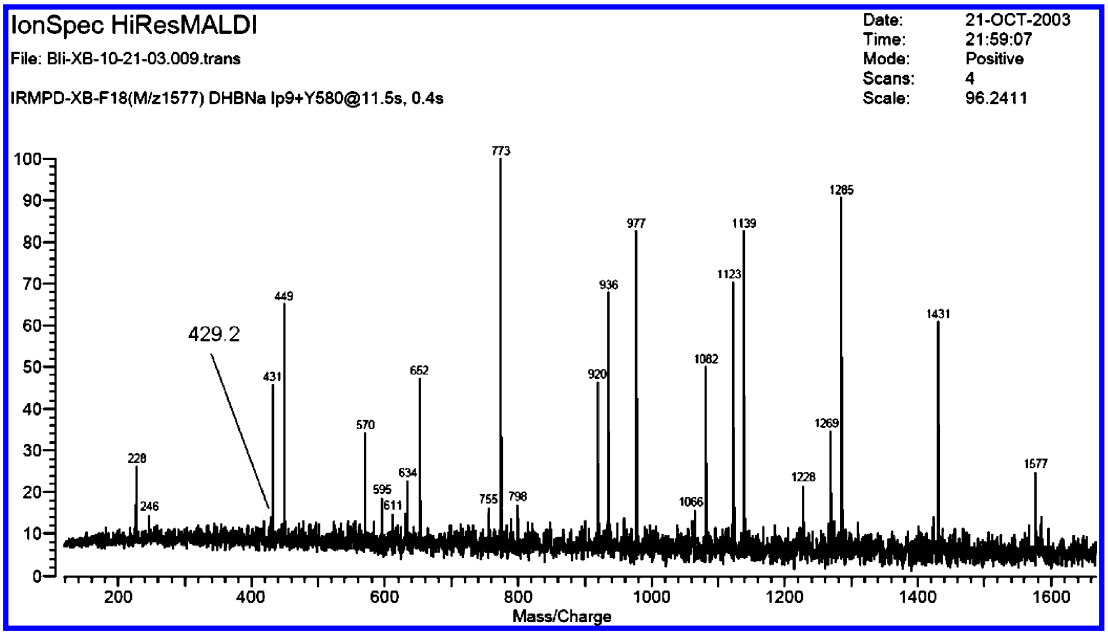

Figure 2. IRMPD FTMS spectrum of an O-glycan. A single IRMPD fragmentation spectrum generally contains all the fragments in a sequence of CID MS^n spectra. Regardless of the type of fragmentation, FTMS spectra have a great many noise peaks, and it is crucial to distinguish small signal peaks from noise. The small peak at 429.2 is two HexNAc's (mass 203.1, not derivatized to alditol) and sodium (mass 23.0). This peak, which would be overlooked by previous thresholding methods, distinguishes the correct topology (on the left in Figure 3) from the other topology.

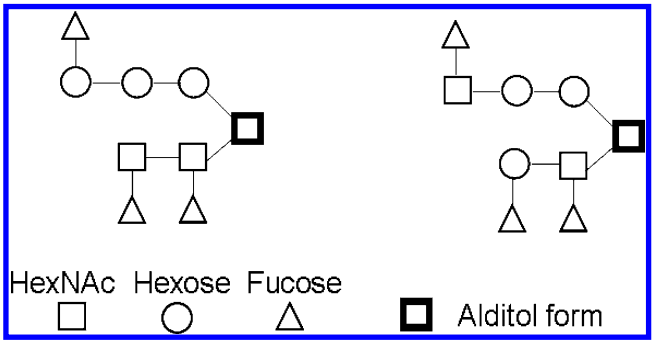

Figure 3. Two candidate topologies (cartoons) for the spectrum shown in Figure 2. Both candidates explain almost all of the large peaks in the spectrum, but the structure on the right cannot explain the peak at *m*/*z* 429.2 corresponding to two HexNAc's without alditol.

In the tentative-peak recalibration method, peaks with sufficiently high *p* values are matched to oligosaccharide masses (including oligosaccharides with single water losses). As mentioned above, oligosaccharide masses are much sparser than peptide masses, so almost all significant peaks will match at most one theoretical mass. A robust statistical regression method is then used to compute a correction curve mapping measured masses to theoretical masses. It is important to use robust regression because tentative peak assignments are often wrong. *CartoonistTwo* uses least-median-of-squares regression[27,28] for outlier rejection, followed by least-squares regression with a quadratic regressor, as shown in Figure 5. We find it interesting that a quadratic regression model greatly outperforms a linear model, which we previously found to be very accurate for time-of-flight (TOF) spectra.[21,26] The difference may be due to a new source of error, such as space-charge effects in FTMS,[25] or to less exact initial calibration.

After recalibration, we obtain measurement errors in the range 0.003−0.005 Da (average absolute value of residuals in each spectrum), which translates to 4−7 ppm. These numbers are comparable with those from other recalibration methods;[24,25] for example, Bruce et al. report errors of 3.6, 7.0, and 5.4 ppm on three spectra using their DeCAL algorithm. Comparing DeCAL with the tentative-peak algorithm, we observe that DeCAL has the advantage that it does not require any knowledge of the chemical species in the spectrum, but the disadvantage that it requires species in multiple charge states (which we do not have here).

2.3.3. Peak Confidence. *CartoonistTwo* models recalibrated mass errors as arising from a normal distribution; it gives each peak a *confidence value* by multiplying the probability density at the peak's mass error by the peak's significance. (The implementation actually adds the logarithms of the probabilities.) Assuming that mass errors and intensities are independent—actually more intense peaks tend to be more accurate—the confidence value gives the probability that the peak indeed represents a glycan fragment. The high mass accuracy of recalibrated FTMS is not usually needed to identify high-intensity glycan peaks, but mass error plays a larger role in determining the overall confidence value for less-intense peaks. It is not uncommon for a spectrum to have a low intensity peak where intensity alone gives a high confidence score and it has mass close to that of a potential glycan, but it has an unreasonably high mass error (after recalibration) so that the total confidence is low. The total confidence probabilities tend to cluster near 100% or 0%. There will typically be 15−60 peaks whose confidence is great than 10%, only 1−2 peaks with confidence probabilities in a range of 10% to 0.1%, and all the rest have scores of less than 0.01%.

CartoonistTwo processes a sequence of MS^n spectra by setting peak significances and recalibrating mass measurements for each spectrum individually, and then taking the union of all the significant peaks observed in all the spectra. A peak observed more than once is given its maximum confidence value. By lumping together all the peaks from the MS^n

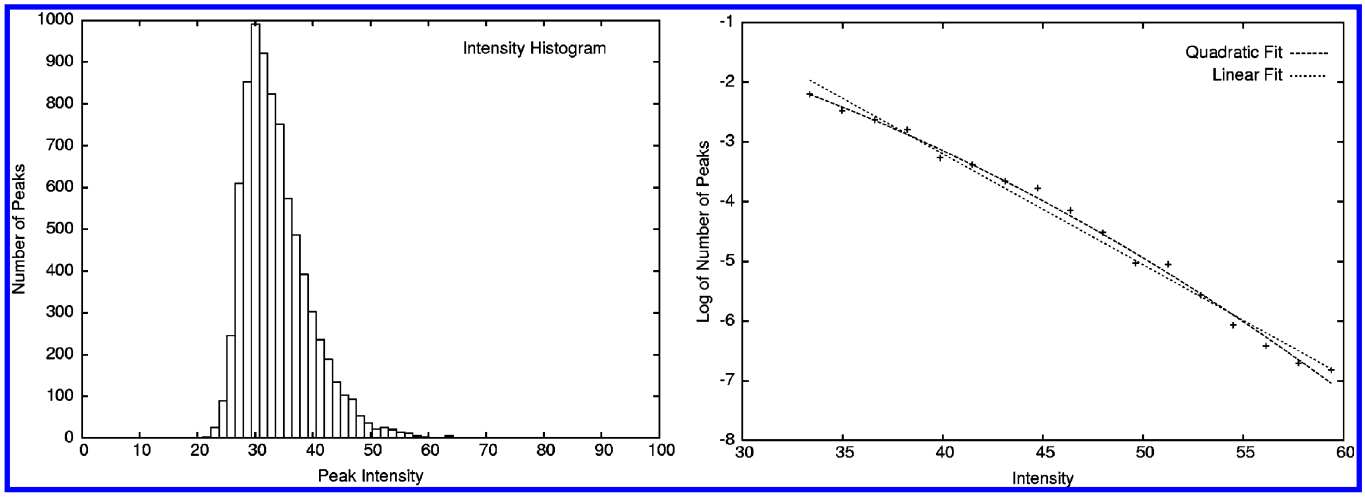

Figure 4. (a) *CartoonistTwo* first computes a histogram of peak intensities. (b) *CartoonistTwo* then graphs (*x*, *y*) pairs for the right tail of the intensity distribution, where *x* is the intensity and *y* is the log of the number of peaks. A fitted curve is used to judge the significance of peak intensities. This method reliably finds small but significant peaks such as the one shown in Figure 2.

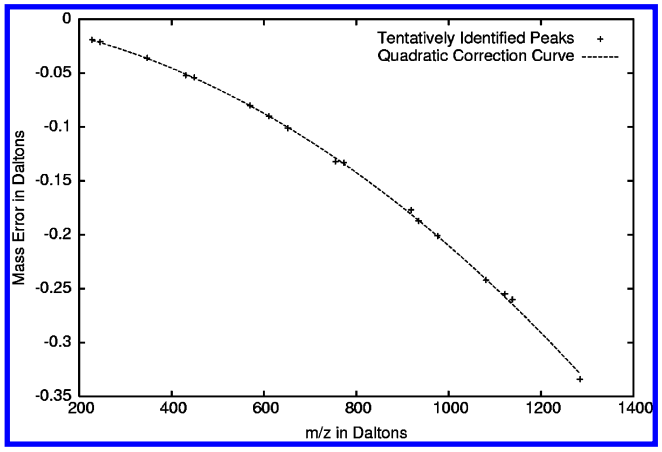

Figure 5. Quadratic regressor is used to recalibrate the *m*/*z* measurements, reducing the average error in each spectrum to 0.003−0.005 Da.

Table 1. Performance of Different Scoring Algorithms against Hand Annotated Spectra[a]

	correct	tie	second	miss	performance
Basic Scorer	7 (7)	27 (27)	2 (0)	3 (0)	0.449 (0.502)
Basic + *Shedding*	9 (9)	25 (24)	2 (1)	3 (0)	0.460 (0.514)
Basic + *Barking*	19 (19)	3 (3)	9 (8)	8 (4)	0.643 (0.716)
Shedding + *Barking*	20 (20)	3 (3)	9 (8)	7 (3)	0.657 (0.730)
Shedding + *Barking (Multiple)*	20 (20)	3 (3)	9 (8)	7 (3)	0.658 (0.732)

[a] Results of the five algorithms on the 39 test spectra. In parentheses are the results excluding the 5 arguable spectra. "Correct" means that the single top-scoring candidate is correct. "Tie" means the correct structure is tied with at least one other topology, "Second" means that the correct structure has the second highest score (possibly tied) and "Miss" means that it was in third or lower position. "Performance" gives a unified accuracy measurement μ, where $1/\mu$ is the expected rank of the correct structure, assuming that topologies with equal scores appear in random order.

sequence, we lose some information—the intensities of the peaks in the individual spectra—but enable the use of the same identification methods for CID MS^n and IRMPD spectra.

2.3.4. Candidate Generation. The third program in the series determines the monosaccharide composition of the parent ion, and then generates and scores topologies. Because most *O*-glycans are made up of only a few different monosaccharides

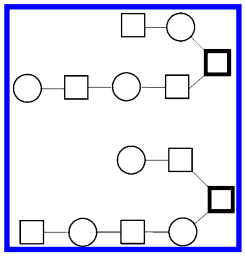

Figure 6. Two glycans may be isospectral, meaning that they have exactly the same set of fragments (although not necessarily with the same multiplicity).

each with a significantly different mass, the monosaccharide composition is usually uniquely determined from the parent-ion mass, and hence sophisticated algorithms such as those devised for peptides[29] are not needed. The generation step next considers *all* possible topologies for the set of monosaccharides, subject to a few biological restrictions on the amount of branching. Specifically the reducing sugar is connected to one or two linear chains of hexoses and HexNAc's and any further branching consists only of fucose sugars. These restrictions limit the number of possible topologies to manageable numbers (less than 10 000) for the spectra used here. Although this includes all the hand annotations for our test spectra, non-conforming topologies are listed in CarbBank,[10] and so different rules might be more appropriate for other applications.

2.3.5. Scoring. A basic scorer, as used by STAT[16] and GLYCH,[17] simply counts the number of spectral peaks (above some threshold intensity) explained by fragments of the candidate (within some mass tolerance). We started with a similarly basic scorer, only rather than using simply zero or one, we used our peak confidences, thereby incorporating both mass error and intensity significance. A peak that is explained twice (that is, by two fragments) scored the same amount as a peak that is explained once. We evaluated this scorer, called *Basic* in Table 1, along with four successively more advanced scorers.

Basic + *Shedding* incorporates the shedding model of glycan fragmentation, as shown in Figure 1. This scorer is the same as *Basic*, but it adds a small constant bonus for each observed fragment with a path of observed peaks to the root of the shedding tree. (The same topology can appear at multiple

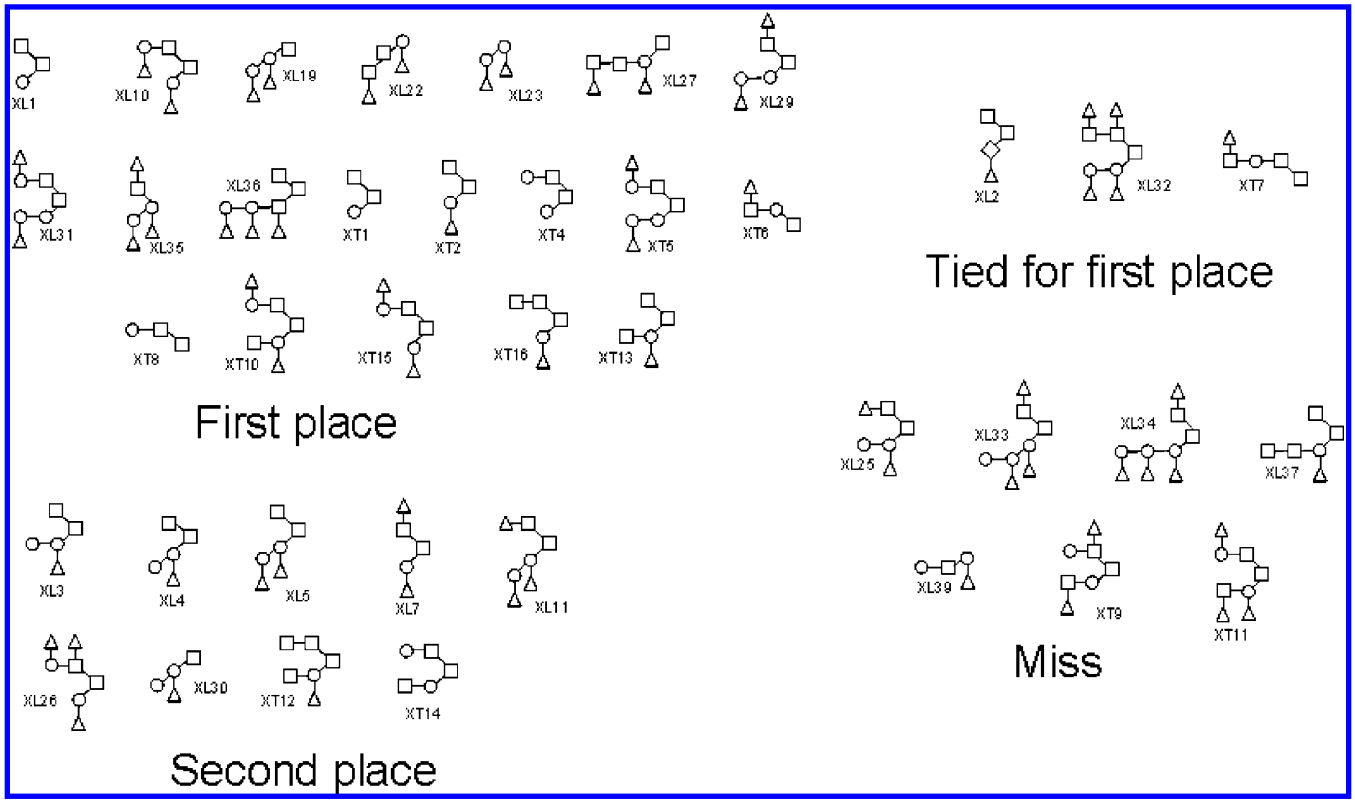

Figure 7. Test set of 39 *O*-glycans. "First place" means that the correct answer was *CartoonistTwo*'s unique top scorer; "Tied for first place" means that the correct answer tied for first with at least one other topology; "Second place" means the correct answer tied for second; and "Miss" is everything else.

nodes of the shedding tree, and a fragment receives the bonus if any of its occurrences has a path to the root.) *Basic* + *Barking* improves the basic scorer in a different way, by penalizing for fragments of the candidate that were not observed—"the dog that did not bark". The penalty was set to a small constant times the number of unobserved fragments. We set the constant very small, so that the unobserved peaks would be used only to break ties between topologies that explained equal numbers of peaks. The question of multiple counting again arises: should a missing peak that corresponds to two different subgraphs of the topology be penalized twice? *Basic* + *Barking* penalized such a missing peak only once. *Shedding* + *Barking* includes both improvements to the basic scorer. Finally, *Shedding* + *Barking* (Multiple) is the same as *Shedding* + *Barking*, but penalizes each missing peak in proportion to the number of subgraphs of that mass that are unobserved.

Results and Discussion

For the experimental evaluation of the five scoring functions, we used 39 sequences of SORI−CID MS^n spectra with known answers. Each sequence included one to five mass spectra, with selection and fragmentation carried out on the dominant peak in the previous spectrum. The known answers were determined by repeated MS experiments and exhaustive manual analysis, and in some cases further validated by Nuclear Magnetic Resonance (NMR).[18] We did not use the newer IRMPD spectra in the test set, as these structures are not yet as well validated.

Results are shown in Table 1 and Figure 7. *CartoonistTwo* (all five scorers) gave low scores to the "known answers" on 5 sequences of spectra, and further examination of the data led us to believe that these manually determined structures were probably wrong. Table 1 lists results for both the original set of 39 sequences and the better-validated set of 34. An important issue in evaluating a glycan scorer is specificity: should we prefer a scorer that often gives the correct topology the top score, but tied with many other top scorers, over a scorer that gives the correct topology the top score less often, but produces fewer ties? We devised a unified performance metric to address this question. If there are no ties, then we assign performance of $\mu = 1/r$ where r is the rank the scorer assigns to the correct topology. So 1.0 means perfect performance, and numbers near zero mean poor performance. If there are ties, then we assume the tied structures appear in random order, and we set $\mu = 1/r$, where r is the expected rank of the correct structure over all possible random orders. Using this metric, we can see that each of the refinements to the basic scorer adds some accuracy. The addition of penalties for predicted peaks not observed (*Barking*) is the largest single improvement, and the improvement offered by the shedding model is small but probably real, as it improved both *Basic* and *Basic* + *Barking*.

Peptide identification algorithms can exploit large databases of spectra and use statistical patterns of fragmentation.[30,31] Since such large databases are not available for glycans, we have followed the more empirical approach of creating reasonable scoring functions and then comparing their performance on a set of test spectra. For example, we make the plausible assumption that every glycosidic bond is equally likely to fragment, and assign an identical small penalty to each missing fragment. As large test sets of spectra become available, we can replace such assumptions with more statistically founded ones.

The final scorer gave a unique top scorer that matched the known answer on 59% of the 34 spectra, about three times as

often as the basic scoring function used up until now. The true performance of the final scorer may actually be somewhat higher, as *CartoonistTwo* has several times caught errors in manual expert annotation on IRMPD spectra (not included in the test set), and it is possible that the curated test set of 34 structures may still include small errors. In several spectra, the correct structure was distinguished from other structures only by the presence or absence of a single small peak, such as the peak at 429.2 in Figure 2.

Cleavage can occur at any glycosidic bond in the molecule, and once one cleavage has occurred, the molecule can continue to fragment to produce additional, smaller oligosaccharides. Thus, in the MS^n spectra, the higher-mass fragments tend to appear in the first fragmentation, or MS^2 spectrum, and most lower-mass fragments do not appear until later rounds of fragmentation and mass measurement (MS^3 and MS^4) as observed previously.[5,19] Fucoses tend to be lost first,[20] and many of the later-round fragments are neither b- nor y-ions as they show multiple losses of leaf monosaccharides. The shedding model hypothesizes that under slow-heating each successive oligosaccharide "sheds" only a single monosaccharide. In terms of the tree in Figure 1, shedding gives a set of oligosaccharides, with each structure having a complete path to the root. The same pattern of connected fragments can also result from a less sequential shedding model, in which an oligosaccharide can lose a two- or three-saccharide component, but enough of the parent ion remains to fill in the gaps in the path. The results support (at least the weaker version of) the shedding model, with 25 of the 39 sequences conforming perfectly, meaning that all the observed fragments had complete paths to the root of shedding tree. In 10 of the remaining 14 sequences, there was only a single discrepancy, that is, only a single observed fragment without a complete path to the root. The most common discrepancy between the shedding model and the observed data was the loss of two fucoses "simultaneously". Although the shedding model is supported by the data, the incorporation into the scorer of a bonus for complete paths to the root did not greatly improve performance, because competing topologies often obtained the same number of complete-path bonuses.

In the context of scoring algorithms, it is worth pointing out that the presence/absence of fragmentation peaks is insufficient to determine the topology of glycans. Figure 6 shows a pair of "isospectral" glycans, that is, distinct topologies with exactly the same set of fragment compositions.

Conclusions

Algorithms have been previously described for the identification of detached glycans from fragmentation spectra, but their usefulness is limited because they typically produce a large number of equally plausible structures.[16] *CartoonistTwo* addresses this problem by introducing a more sophisticated scoring function that makes finer discriminations between structures. Our experiments show that scorers that take into account fragments that do not appear in the spectrum offer a substantial improvement over simpler scoring functions. To use this improvement, however, it is crucial that the processing pipeline accurately distinguish low-intensity signal peaks from background noise. With its use of a noise model and m/z recalibration, *CartoonistTwo* can effectively cull the list of possible structures.

We have presented results only for *O*-glycans, but we could easily adapt *CartoonistTwo* to handle *N*-glycans or free oligosaccharides, with changes necessary only in the candidate generation step. In fact *O*-glycans are more difficult than the more well-studied *N*-glycans, because they have fewer known biological constraints. The current version of *CartoonistTwo* determines only the topology of an *O*-glycan, but it has been shown that differing linkages can produce characteristic fragmentation patterns,[18] even in the absence of cross-ring fragmentation, so in future work we plan to explore the possibility of computing linkage information as well.

Acknowledgment. D.G. was supported by NIH Grant No. R01GM074128-01 from the NIGMS and resources from the Consortium for Functional Glycomics funded by grants from the NIGMS (GM62116) and the NCRR. C.L. was supported by NIH Grant No. R01GM049077.

References

(1) Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta* **1999**, *1473*, 4−8.
(2) Taylor, M. E.; Drickamer, K. *Introduction to Glycobiology*; Oxford University Press: New York, 2003.
(3) Chou, H. H.; Hayakawa, T.; Diaz, S.; Krings, M.; Indriati, E.; Leakey, M.; Paabo, S.; Satta, Y.; Takahata, N.; Varki, A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11736−11741.
(4) Chou, H. H.; Takematsu, H.; Diaz, S.; Iber, J.; Nickerson, E.; Wright, K. L.; Muchmore, E. A.; Nelson, D. L.; Warren, S. T.; Varki, A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11751−11756.
(5) Zhang, J.; Schubothe, K.; Li, B.; Russell, S.; Lebrilla, C. B. *Anal. Chem.* **2005**, *77*, 208−214.
(6) Joshi, H. J.; Harrison, M. J.; Schulz, B. L.; Cooper, C. A.; Packer, N. H.; Karlsson, N. G. *Proteomics* **2004**, *4*, 1650−1664.
(7) Cooper, C. A.; Gasteiger, E.; Packer, N. H. *Proteomics* **2001**, *1*, 340−349.
(8) Lohmann, K. K.; von der Lieth, C. W. *Nucleic Acids Res.* **2004**, *32*, W261−266.
(9) Cooper, C. A.; Joshi, H. J.; Harrison, M. J.; Wilkins, M. R.; Packer, N. H. *Nucleic Acids Res.* **2003**, *31*, 511−513.
(10) Doubet, S.; Albersheim, P. *Glycobiology* **1992**, *2*, 505.
(11) Loss, A.; Bunsmann, P.; Bohne, A.; Loss, A.; Schwarzer, E.; Lang, E.; von der Lieth, C.-W. *Nucleic Acids Res.* **2002**, *30*, 405−408.
(12) Eng, J. K.; McCormack, A. L.; Yates, J. R. I. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(13) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(14) Ethier, M.; Saba, J. A.; Ens, W.; Standing, K. G.; Perreault, H. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1743−1754.
(15) Ethier, M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2713−2720.
(16) Gaucher, S. P.; Morrow, J.; Leary, J. A. *Anal. Chem.* **2000**, *72*, 2331−2336.
(17) Tang, H.; Mechref, Y.; Novotny, M. V. Detroit, Michigan 2005; i431−i439.
(18) Tseng, K.; Hedrick, J. L.; Lebrilla, C. B. *Anal. Chem.* **1999**, *71*, 3747−3754.
(19) Xie, Y.; Lebrilla, C. B. *Anal. Chem.* **2003**, *75*, 1590−1598.
(20) Zaia, J. *Mass. Spectrom. Rev.* **2004**, *23*, 161−227.
(21) Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A. *Proteomics* **2005**, *5*, 865−875.
(22) Masselon, C.; Anderson, G. A.; Harkewicz, R.; Bruce, J. E.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000**, *72*, 1918−1924.
(23) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320−332.
(24) Flora, J. W.; Muddiman, D. C. *Anal. Chem.* **2001**, *73*, 3305−3311.
(25) Bruce, J. E.; Anderson, G. A.; Brands, M. D.; Pasa-Tolic, L.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 416−421.
(26) Bern, M.; Goldberg, D. *J. Comput. Chem.* **2006**, *13*, 364−378.
(27) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley: New York, 1987.
(28) Fischler, M. A.; Bolles, R. C. *CACM* **1981**, *24*, 381−395.
(29) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703−714.
(30) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 1155−1163.
(31) Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75*, 435−444.

PR060035J