COMPLEXITIES AND ALGORITHMS FOR GLYCAN STRUCTURE SEQUENCING USING TANDEM MASS SPECTROMETRY*

BAOZHEN SHAN, BIN MA AND KAIZHONG ZHANG

Department of Computer Science
University of Western Ontario
London, Ontario, Canada N6A 5B7
E-mail: bxshan, bma, kzhang@csd.uwo.ca

GILLES LAJOIE

Department of Biochemistry
University of Western Ontario
London, Ontario, Canada N6A 5B7
E-mail: glajoie@uwo.ca

Determining glycan structures is vital to comprehend cell-matrix, cell-cell, and even intracellular biological events. Glycan structure sequencing, which is to determine the primary structure of a glycan using MS/MS spectrometry, remains one of the most important tasks in proteomics. Analogous to the peptide *de novo* sequencing, the glycan *de novo* sequencing is to determine the structure without the aid of a known glycan database. We show in this paper that glycan *de novo* sequencing is NP-hard. We then provide a heuristic algorithm and develop a software program to solve the problem in practical cases. Experiments on real MS/MS data of glycopeptides demonstrate that our heuristic algorithm gives satisfactory results on practical data.

1. Introduction

The carbohydrates of glycoproteins and glycolipids are commonly referred as glycans. The glycan moieties cover a range of diverse biological functions. The rapid progress in proteomics has generated an increased interest in the full characterization of glycans [1].

A glycan is assembled from simple sugars by removal of water during the linkage of simple sugars. In terms of mass, there are eight usual types of simple sugars. There are two types of glycans, O-linked and N-linked glycans. The primary structure of a glycan is characterized by its two dimensional linkages (Figure 1a), which can be represented by labelled trees with node labels representing the type of simple sugars (Figure 1b).

Glycan structure sequencing is to determine primary structures of glycans using tandem mass spectrometry [6]. In an MS/MS experiment, a glycan will fragment into different

*This research was undertaken, in part, thanks to funding from NSERC, PREA, the Canada Research Chairs program, and NSF of China 60553001. BM's work was partially done when he visited the Center for Advanced Study at Tsinghua University.

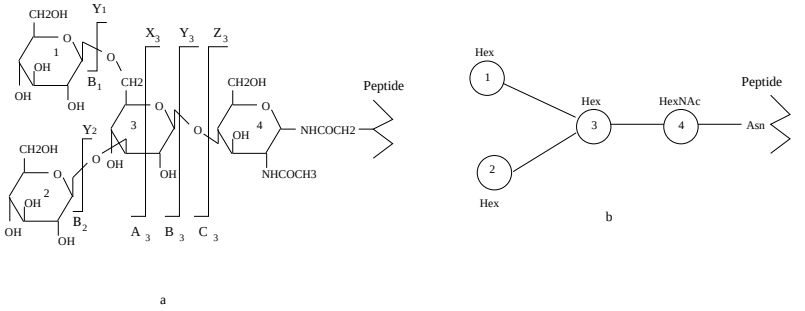

Figure 1. (a) Four simple sugars are linked together to form a glycan on a peptide. The long verticle lines indicate some possible fragmentations of the glycan in an MS/MS experiment. A_i, B_i, C_i, X_i, Y_i, and Z_i are the names of the resulted fragment ions. (b) The abstract tree representation of the glycan in (a).

fragment ions at any of the several labeled locations. Modified Domon and Costello's [7] nomenclature for glycan fragmentation is used in this paper (Figure 1a). There are three types of fragmentation, generating six types of ions A, B, C, X, Y and Z. A, B, and C types of ions correspond to subtrees. Whereas, X, Y, and Z types of ions correspond to the precursor (the whole structure including the peptide) minus subtrees. Usually, Y and B ions dominate in the spectrum [18]. So, for simplicity, we consider only Y and B ions to demonstrate our models.

The MS/MS spectrum of a glycan consists of many peaks, each of which is presumably generated by many copies of one fragment ion. The position of the peak indicates the mass to charge ratio of the corresponding fragment ion, and the intensity of the peak indicates the relative abundance of the fragment ion. Consequently, different glycans usually produce different MS/MS spectra. It is thus possible, and now a common practice, to use the spectrum of a glycan to determine its sequence structure.

One method to interpret the MS/MS spectrum involves searching the glycan database for a glycan, of which the theoretical spectrum matches the observed spectrum well. However, the determination of novel glycan structures requires *de novo* sequencing, which computes the primary structure directly from the spectrum without the help of a glycan database. Because the glycan database is rather incomplete, currently *de novo* sequencing is more important in these two approaches.

Tandem mass spectrometry has perviously been used to sequence peptides, where the linear amino acid sequence of a peptide is derived from the MS/MS spectrum (see [3,13,2] for a few examples). Glycan *de novo* sequencing in some sense is more difficult as it tries to derive a tree structure instead of a linear sequence. There have been attempts to solve glycan *de novo* sequencing problem [4,5,8,9,10,14,15,16]. Gaucher *et al.* [9] reported a glycan topology analysis tool, STAT, which generates all possible glycan primary structures. Obviously, the number of possible glycan structures grow exponentially. Consequently, STAT is only feasible for small glycans containing up to ten monosaccharides.

Similar to spectrum graph approach of peptide *de novo* sequencing [2], relationship between peaks in the spectrum is used to reduce the computation in glycan *de novo* sequenc-

ing. Mizuno et al. [14] first employed the relationship for compositional analysis. Ethier et al. [8] further built relationship tree for glycan *de novo* sequencing. Relationship tree approach has some difficulties to deal with. First, it requires very high quality of spectrum. Second, the restriction of certain structural topology makes it inappropriate for general glycan structures.

Shan *et al*. [15] reported a heuristic algorithm for glycan *de novo* sequencing from the MS/MS spectra of glycopeptides. The algorithm first generates many acceptable small subtrees, which are then joined together in a repetitive process to obtain larger and larger suboptimal subtrees until reaching the desired mass. At each size of the subtree, only limited number of subtrees are kept for later use. Experiments on real MS/MS data showed that the heuristic algorithm can determine glycan structures.

Tang *et al*. [16] reported a dynamic programming approach to determine glycan structures by recording the k best solutions at each iteration step (by default $k = 200$). The algorithm prefers linear structure to branching structure. Another algorithm called Cartoonist [10] has been created to interpret spectra of N-glycans. For each peak, Cartoonist computes all plausible structures with scores indicating the probability of correctness.

The classical methods for the characterization of glycoproteins by mass spectrometry were to cleave glycans with enzymes and then analyze the structures of the released glycans. Therefore, most of reported algorithms focus on interpreting MS/MS spectra of released glycans [4,5,8,9,10,14,16]. Recently, biochemists began to analyze glycopeptides derived from trypsin digestion of glycoproteins directly [18,12]. Therefore, there is a need to provide computational tools to assist the analyses. The algorithm in [15] was designed for both released glycan and glycopeptide. However, it only worked for the N-linked glycopeptides. The other type, O-linked glycopeptides often contains more than one glycan moieties.

Before this research, no software has been reported for interpreting spectrum of glycopeptide with multiple glycan moieties. In addition, the computational complexity of glycan *de novo* sequencing has remained unknown. The paper aims to solve these problems. More specifically, the contributions of this paper are the following:

(1) A polynomial time algorithm is provided under a simple model of glycan *de novo* sequencing.
(2) A more realistic model of glycan *de novo* sequencing is proved to be NP-hard.
(3) A new heuristic algorithm for the glycan *de novo* sequencing is introduced.
(4) A software program, GlycoMaster, is developed. Experiments on real data demonstrate our method works very well in practice. The software is available at http://bif.csd.uwo.ca/glycomaster.

2. Modelling Glycan *De Novo* Sequencing Problem

In the MS/MS spectrum of a glycopeptide, peaks of Y and B ions are located at separate regions because Y fragments contain peptide part which usually has relatively large mass. Let Σ be the alphabet of simple sugars. A glycan tree T is an unordered rooted tree with

bounded degree whose nodes are labelled by letters from Σ. The degree of glycan trees is bounded by four because each sugar has at most five linkages. The root of T is linked to a peptide $P = a_1 \dots a_k$ where a_i is from an amino acid alphabet Σ_a.

We assume that there is a post order numbering of the nodes in T. We use t_i to represent the ith node in T and $T[i]$ to represent the subtree rooted at t_i. When there is no confusion, we also use t_i to represent the sugar located at node t_i. Let $|T'|$ denote the size of a tree T'. Because of the post order numbering, for each $1 \leq i \leq |T|$, there is an $\hat{i} = i - |T[i]| + 1$ such that $t_{\hat{i}}, t_{\hat{i}+1}, \dots, t_i$ represent all the nodes in $T[i]$.

For a sugar $g \in \Sigma$, we use $\|g\|$ to denote its mass. For an amino acid $a \in \Sigma_a$, we use $\|a\|$ to denote its mass. Let $\|T\| = \sum_{i=1}^{|T|} \|t_i\|$ and $\|P\| = \sum_{i=1}^{|P|} \|a_i\|$, then the actual mass, M, of the precursor ion of tree T linked with peptide P is $\|T\| + \|P\| + 18$ because of an extra H_2O in the peptide. For each subtree $T[i]$, let B_i represent the B-ion associated with $T[i]$ and Y_i represent the Y-ion associated with T linked with P and subtree $T[i]$ removed. Let $b_i = \sum_{k=\hat{i}}^{i} \|t_k\|$ and $y_i = M - b_i$, the actual mass of B_i is $b_i + 1$ (because of a proton added) and the mass of Y_i is y_i. We use $Y_{i_1,\dots,i_k}$, where $T[i_1], \dots, T[i_k]$ are non-overlapping subtrees, to denote the Y-ion associated with T linked with P and subtrees $T[i_1], \dots, T[i_k]$ removed. Let $y_{i_1,\dots,i_k} = M - (\sum_{l=1}^{k} b_{i_l})$, then the mass of $Y_{i_1,\dots,i_k}$ is $y_{i_1,\dots,i_k}$.

For simplicity, in this section, we only consider B-ions and Y-ions of the form of B_i and Y_i.

Let $\mathcal{M} = \{(m_i, h_i)\}$ be a spectrum of a glycan, where m_i is the mass and h_i is the intensity of the peak. For each mass value m, according to the intensity of the peaks around m, a score function $g(m)$ can be defined. Let T be a glycan tree, then the score of T, $S(T)$ is defined to be the summation of $g(m)$ for all the mass values m of the fragment ions of T. And the glycan structure *de novo* sequencing problem is then to find a tree structure T such that the mass of T is equal to a given value M', and $S(T)$ is maximized.

Notice that several different fragment ions of T may give the same mass value m. In such a case, whether the $g(m)$ score is counted several times or only once changes the definition of $S(T)$. Consequently, the difficulties of the glycan structure *de novo* sequencing problem under these two definitions are very different. The following two sections discuss these two definitions.

2.1. *A Simple Model*

Given a mass spectrum $\mathcal{M}$, there are different ways to define the score function $g(m)$. In this paper we simply assume $g(m)$ is given. Let $M' = M - \|P\| - 18$, where M is the precursor mass and P is the peptide. Let T be a glycan tree. The score of T is defined as follows: $S(T) = \sum_{i=1}^{|T|} g(b_i + 1) + g(y_i)$. Furthermore, because $y_i = M - b_i$, we denote $f(m) = g(m + 1) + g(M - m)$. Then the score of a tree T becomes

$$S(T) = \sum_{i=1}^{|T|} f(b_i). \tag{1}$$

It turns out that under such definition, we can in polynomial time calculat the opti-

mal tree T such that $\|T\| = M'$ and $S(T)$ is maximized. Two dynamic programming algorithms can be designed based on Lemma 2.1 and 2.2, respectively. The proofs of the lemmas are omitted here and will be provided in the full version of the paper. It is not hard to prove that the time complexities of the two algorithms are $O(M^4)$ and $O(M^2)$, respectively.

Lemma 2.1. *Let $S(m)$ be the maximum score of a glycan tree with mass m, then*

$$S(m) = \max_{\substack{g \in \Sigma \\ 0 \le m_1 \le m_2 \le m_3 \le m_4 \\ m_4 = m - \|g\| - m_1 - m_2 - m_3}} f(m) + S(m_1) + S(m_2) + S(m_3) + S(m_4)$$

Lemma 2.2. *Let $S(m)$ be the maximum score of a glycan tree with mass m and $S_2(m)$ be the maximum score of a glycan forest with at most two glycan trees and mass m, then*

$$S(m) = \max_{g\in\Sigma; 0\le m_1\le m-\|g\|-m_1} f(m) + S_2(m_1) + S_2(m - \|g\| - m_1)$$
$$S_2(m) = \max_{0\le m_1\le m-m_1} S(m_1) + S(m - m_1)$$

The simple model in this section works in practice if the spectrum has good quality and the structure is relatively simple. The problem with this model is that if there are two different subtrees $T[i]$ and $T[j]$ with the same mass, $m(T[i]) = m(T[j])$, then in the score for T, $S(T)$, this mass will be used twice. Since these two ions generate the same peak in the spectrum, we should only use the peak once.

2.2. *A More Realistic Model*

In order to avoid repeatedly using peaks, the score function defined in (1) needs to be modified as follows.

Let $\Gamma(T) = \{b_i, y_i \mid 1 \le i \le |T|\}$. Define $S(T) = \sum_{m\in\Gamma(T)} g(m)$. Because $\Gamma(T)$ is a set, the definition here only uses each mass value once, even if there are multiple ions giving the same value. Further, the existence of the peptide separates Y and B ions in the MS/MS spectrum of a glycopeptide. If we let $\Delta(T) = \{b_i \mid 1 \le i \le |T|\}$. Then $S(T)$ can be rewritten as

$$S(T) = \sum_{m\in\Delta(T)} f(m). \tag{2}$$

With this model, the approach of Lemma 2.2 no longer works. In fact, the complexity of computing the optimal solution under this model becomes NP-hard.

3. NP-hardness of Glycan *De Novo* Sequencing

We will reduce Exact Cover by 3-Sets [11] to glycan structure sequencing problem.

Exact Cover by 3-Sets

INSTANCE: A finite set $E = \{e_1, e_2, \ldots, e_n\}$ where $n = 3q$ and a collection S of 3-element subsets of E.

QUESTION: Does S contain an exact cover for E, that is, a subcollection $S' \subseteq S$ such that every element of E occurs in exactly one member of S'?

Lemma 3.1. *There are n positive integers $z_1 < z_2 \ldots < z_n$ such that*

(1) $z_i + z_j = z_{i'} + z_{j'}$ *implies* $\{i, j\} = \{i', j'\}$.
(2) $z_i + z_j + z_k = z_{i'} + z_{j'} + z_{k'}$ *implies* $\{i, j, k\} = \{i', j', k'\}$.
(3) $z_i \leq poly(n)$.

Proof. We determine $z_1, \ldots, z_n$ inductively. The first three integers are $z_1 = 1$, $z_2 = 2$, and $z_3 = 3$. Suppose that we already determined $z_1, \ldots, z_{k-1}$ satisfying conditions 1 and 2, we now prove that z_k can be found in the range of $z_{k-1} + 1$ and $z_{k-1} + n^5$.

Consider the following two equations where $1 \leq i, j, l, i', j' \leq k - 1$.

- $x + z_{i'} = z_i + z_j$
- $x + z_{i'} + z_{j'} = z_i + z_j + z_l$

There are in total less than $\frac{n^3}{2} + \frac{n^5}{5!}$ equations above, each has only one solution of x. Therefore, in the range of $z_{k-1}+1$ and $z_{k-1}+n^5$, there is an integer that is not the solution of any of the equations. The integer can be trivially found in polynomial time and is used as z_k.

From the construction, we have $z_1 < \ldots < z_n = O(n^6)$. □

Lemma 3.2. *There are n positive integers $z_1, \ldots, z_n$ satisfying the conditions in Lemma 3.1 and*

(4) if $i \neq j$, $|z_i - z_j| \geq n^6 + 2$.
(5) if $\{i, j\} \neq \{i', j'\}$, $|z_i + z_j - z_{i'} - z_{j'}| \geq n^6 + 2$.
(6) if $\{i, j, k\} \neq \{i', j', k'\}$, $|z_i + z_j + z_k - z_{i'} - z_{j'} - z_{k'}| \geq n^6 + 2$.

Proof. Mutiply each integer determined in Lemma 3.1 by $n^6 + 2$. This will not violate any conditions in Lemma 3.1. □

Let Σ contain only one letter g, and $m(g) = 1$.

Theorem 3.1. *The glycan structure sequencing problem is NP-hard under the realistic tree scoring model and an arbitrary mass scoring scheme.*

Proof. Due to page limit, we only give the reduction and the idea of the proof. More details of the proof will be provided in the full version of the paper.

Given an instance of the exact cover by 3-sets with $E = \{e_1, e_2, \ldots, e_n\}$ and $S = \{s_1, \ldots, s_q\}$ where $n = 3q$ and $s_l = \{e_i, e_j, e_k\}$.

Our idea is to design n subtrees, $T_1, \ldots, T_n$, each corresponding to an e_i. By carefully designing the spectrum, i.e., assigning values of $f(m)$ at different m, we can ensure that when there is an exact cover, then the optimal solution will be a tree like Figure 2, where

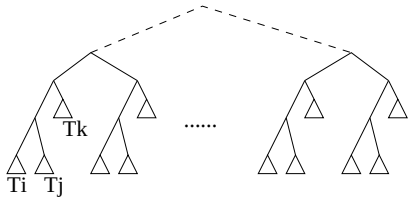

Figure 2. Optimal tree of our construction in NP-hardness proof.

each three-subtree group corresponds to a 3-set. And the score of the tree is equal to n. However, if there is no exact cover, then the score of the tree is smaller than n.

Let $z_1, \ldots, z_n$ be as in Lemma 3.2 and $N = n \times \max z_i$. Let $M = 4 \times N$. Let $x_i = M + z_i$. Then $x_1, \ldots, x_n$ also satisfy Lemma 3.2.

Let e_i correspond to x_i. Let $s_l = \{e_i, e_j, e_k\}$ correspond to $z_l = x_i + x_j + x_k + 2$. Let $X = \{x_i\}$. Let $Y = \{y_k \mid y_k = x_i + x_j + 1\}$. Let $Z = \{z_l\}$.

Let $D_1 = \{|x_i - x_j| - 1\}$, $D_2 = \{|x_i + x_j - x_{i'} - x_{j'}| - 1\}$, and $D_3 = \{|x_i + x_j + x_k - x_{i'} - x_{j'} - x_{k'}| - 1\}$. Let $D = D_1 \cup D_2 \cup D_3$, then $|D| < n^6$ and for any $d \in D$, $n^6 < d < N$.

In the spectrum, let

$$f(m) = \begin{cases} -1, & m \in D \\ 0, & 1 \le m \le M - N \text{ and } m \notin D \\ 1, & m \in X \\ -1, & M - N < m \le M + N \text{ and } m \notin X \\ 0, & m \in Y \\ -1, & 2M < m \le 2M + N \text{ and } m \notin Y \\ 0, & m \in Z \\ -1, & 3M < m \le 3M + N \text{ and } m \notin Z \\ 0, & 3tM < m \le 3tM + N,\ 2 \le t \le n/3 \\ -1, & \text{else} \end{cases}$$

Let the total mass be $\sum x_i + n - 1 = nM + \sum z_i + n - 1 < nM + N$.

Note that the score function $f()$ can be computed in $poly(n)$ time. Because of the construction of $f()$, the following property can be proved.

Claim 3.1 There is a subtree T_i for each x_i, such that $m(T_i) = x_i$ and $f(m(T')) = 0$ for each subtree T' of T_i.

If there is an exact cover $S' \subset S$, for every $s_i = \{e_i, e_j, e_k\} \in S'$, we can link T_i, T_j, and T_k together as in Figure 2. This will give us a total score n because all $f(m) = 1$ for $m \in X$ are included, while no $f(m) = -1$ is used.

On the other hand, if there is a tree with score n, then all $f(m)$ for $m \in X$ are included and none of $f(m) = -1$ is used. This can be used to prove that in the optimal tree, there

is one and only one subtree with mass x_i. and for each such subtree, its parent or grand parent is of mass $x_i + x_j + x_k + 2$ where $\{e_i, e_j, e_k\}$ is in S. This means that we have an exact cover. □

4. An Algorithm for Glycan *De Novo* Sequencing

In previous sections, we showed that Glycan *De Novo* sequencing problem has polynomial algorithm to find optimal solution with the simple model. However, with more realistic model it is NP-hard. In this section, we extend the polynomial time algorithm of the simple model to a heuristic algorithm of the realistic model.

4.1. *A Heuristic Algorithm*

A glycan forest is represented as $f =< t_1, \ldots, t_n >$ where t_i is a glycan tree and n is called the degree of f. We use t_0 denotes an empty tree. We use $F(m)$ to represent a set of glycan forests with mass m.

$$F(m) = \{f \mid f \text{ is a forest such that } m(f) = m\}$$

We use $g \oplus f$ to represent the glycan tree rooted at g and each tree of f is a child of g. Let $f \ominus < t_i, t_j, t_k, t_l >$ represent the glycan forest resulting from removing t_i, t_j, t_k, t_l from f.

Given a $g \in \Sigma$ and a glycan forest f, we use $g \otimes f$ to represent a set of glycan forests generated by g and $f =< t_1, \ldots, t_n >$.

$$g \otimes f = \{< \ g \oplus < t_i, t_j, t_k, t_l >, f \ominus < t_i, t_j, t_k, t_l > \ > \mid 0 \leq i \leq j \leq k \leq l \leq n\}$$

In our algorithm, the score of a forest is computed by scoring a set of masses of fragments which include Y-fragments, B-fragments and internal fragments (B-fragment losing one or more B-fragments). Fragments with same mass are used only once, because these fragments generate the same peak in the spectrum.

We now give a high level description of our heuristic algorithm. For each mass m we maintain a fixed number of forests with mass m and high scores in $F(m)$. We can compute $F(m)$ in two steps

(1) Compute a set of candidate forests $F_c(m) = \cup_{g \in \Sigma} \cup_{f \in F(m-||g||)} \ g \otimes f$
(2) Compute $F(m)$ by evaluating the score for each of the forests in $F_c(m)$ and remove those with low scores.

Let M' be the desired mass of the glycan forest. Once $F(M')$ is computed, the solution is chosen from forests in $F(M')$ with maximum score.

Algorithm:

Input: Σ (simple sugars), M' (mass of the glycan) and $\mathcal{M}$ (a set of mass peaks)
Output: candidates of sequence structures with scores
(1) **for** $m = 1$ **to** M'
(2) $F_c(m) = \emptyset$
(3) **for** each $g \in \Sigma$

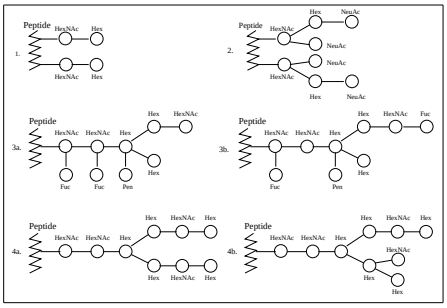

Figure 3. Sample structures computed by GlycoMaster

(4) **for** each $f \in F(m - ||g||)$
(5) **for** each (i, j, k, l) s.t. $(t_i, t_j, t_k, t_l \in f)$
(6) $F_c(m) = F_c(m) \cup < g \oplus < t_i, t_j, t_k, t_l >, f \ominus < t_i, t_j, t_k, t_l > >$
(7) Scoring forests in $F_c(m)$ and put top $|F|$ in $F(m)$
(8) **return** $F(M')$

The time complexity is $O(d^4|\Sigma| \times |F| \times M')$, where d is the degree of forest.

4.2. *Experiments and Discussion*

Based on the above algorithm, a software program, GlycoMaster, has been developed for glycan *de novo* sequencing. We choose $|F| = 1000$ in the software implementation of the algorithm.

The software was tested using twenty MS/MS spectra of glycopeptides. The samples were derived from the cationic peanut peroxidase and rat bone osteopontin after tryptic digestion. The MS/MS spectra of the samples were obtained by using a Q-TOF2 in the positive ion ESI MS/MS mode with borosilicate nano tips. The correctness of the automated interpretation was evaluated by comparing with the structure determined by manual analysis [12,19]. Because most algorithms [4,5,8,9,14,10,16] were designed for MS/MS spectra of glycans only and therefore cannot handle glycopeptides data we obtained, here we only demonstrate that the algorithm in this paper improved our previous algorithm in [15].

The performance of our algorithm is shown in the following table.The compositions of all samples were correctly computed. The structures of nineteen out of twenty spectra are the same as deduced by manual interpretation. Two of them are O-linked glycopeptides with two glycan moieties linked to the peptide (1 and 2 in Figure 3).

Algorithm	Correct Structures	Wrong Structures	Partially Wrong
[15]	16/20	2/20	2/20
This work	19/20	0/20	1/20

Our algorithm also works for the MS/MS data of released glycans. As future work the algorithm will be tested with MS/MS data of released glycans. Currently, no public MS/MS data of glycans are available to us.

References

1. K. Aebersold and M. Mann. Mass spectrometry-based protemics. *Nature*, 422:198-207, 2003.
2. V. Bafna and N. Edwards. On de novo interpretation of tandem mass spectra for peptide identification. *RECOMB 2003*, 9-18, Berlin, Germany, 2003.
3. Chen, T., Kao, M-Y., Tepel, M., Rush J., and Church, G. 2001. A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comp. Biology* 8(3), 325-337.
4. C.A. Cooper, E. Gasteiger, and N.H. Packer. GlycoMod - A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1:340-349, 2000.
5. C.A. Cooper, H. Joshi, M. Harrison, M. Wilkins, and N Packer: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, 31:511-513, 2003.
6. A. Dell and H. R. Morris. Glycoprotein Structure Determination by Mass Spectrometry. *Science*, 291:2351-2356, 2001.
7. B. Domon and CE. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.*, 5:397-409, 1988.
8. M. Ethier, *et al.*. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from tandem mass spectrometry.*Rapid Commun. Mass Spectrom.*, 17:2713-2720, 2003.
9. S.P. Gaucher, J. Morrow and J.A. Leary. A saccharide topology analysis tool used in combination with tandem mass spectrometry *Anal. Chem.*, 72:2231-2236, 2000.
10. D. Goldberg, M. Sutton-Smith, J. Paulson and Anne Dell. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics*, 5:865-875, 2005.
11. Michael R. Garey and David S. Johnson. Computers and intractability: a guide to the theory of NP-completeness. pp221, W.H. Freeman and Company, San Franciso 1979.
12. K. Keykhosravani, A. Doherty-Kirby, C. Zhang, A. Goldberg, G.K. Hunter and G. Lajoie. Comprehensive identification of posttranslational modifications in rat bone osteopontin by mass spectrometry. *Biochemistry*, 44:6990-7003, 2005.
13. B. Ma et. al. PEAKS: Powerful Software for Peptide De Novo Sequencing by Tandem Mass Spectrometry. *Rapid Comm. in Mass Spectrom.* 17(20):2337-2342. 2003.
14. Y. Mizuno and T. Sasagawa. An automated interpretation of MALDI/TOF postsource decay spectra of oligosaccharides. 1. Automated Peak Assignment. *Anal. Chem.*, 71:4764-4771, 1999.
15. B. Shan, K. Zhang, B. Ma, C. Zhang and G. Lajorie. An Algorithm for Determining Glycan Structures from MS/MS Spectra. *Proceedings of ICBA 2004*, Florida, USA, 2004, in Advanced in bioinformatics and its applications, 414-425, World Scicntific, 2004.
16. H. Tang, Y. Mechref, and M.V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 21:i431-i439, 2005.
17. M.E. Taylor and K. Drickamer. Introduction to Glycobiology. Oxford University Press, 2003.
18. J. Zala. Mass Spectrometry of Oligosaccharides. *Mass Spectrometry Reviews*, 23161-227, 2004.
19. C. Zhang, A. Doherty-Kirby and G. Lajoie. Investigation of Cationic Peanut Peroxidase Glycans by Electrospray Ionization Mass Spectrometry. *Phytochemistry*, 65:1575-88, 2004.