Systems biology

Glycan family analysis for deducing N-glycan topology from single MS

David Goldberg1,*, Marshall Bern1, Simon J. North2, Stuart M. Haslam2 and Anne Dell2

1Palo Alto Research Center, 3333 Coyote Hill Rd, Palo Alto CA 94304, USA and 2Division of Molecular Biosciences, Faculty of Natural Sciences, South Kensington Campus, Biochemistry Building, Imperial College, London SW7 2AZ, UK

Received on September 11, 2008; revised on November 9, 2008; accepted on December 06, 2008

ABSTRACT

Motivation: In the past few years, mass spectrometry (MS) has emerged as the premier tool for identification and quantification of biological molecules such as peptides and glycans. There are two basic strategies: single-MS, which uses a single round of mass analysis, and MS/MS (or higher order MS*), which adds one or more additional rounds of mass analysis, interspersed with fragmentation steps. Single-MS offers higher throughput, broader mass coverage and more direct quantitation, but generally much weaker identification. Single-MS, however, does work fairly well for the case of N-glycan identification, which are more constrained than other biological polymers. We previously demonstrated single-MS identification of N-glycans to the level of ‘cartoons’ (monosaccharide composition and topology) by a system that incorporates an expert’s detailed knowledge of the biological sample. In this article, we explore the possibility of ab initio single-MS N-glycan identification, with the goal of extending single-MS, or primarily-single-MS, identification to non-expert users, novel conditions and unstudied tissues.

Results: We propose and test three cartoon-assignment algorithms that make inferences informed by biological knowledge about glycan synthesis. To test the algorithms, we used 71 single-MS spectra from a variety of tissues and organisms, containing more than 2800 manually annotated peaks. The most successful of the algorithms computes the most richly connected subgraph within a ‘cartoon graph’. This algorithm uniquely assigns the correct cartoon to more than half of the peaks in 41 out of the 71 spectra.

Contact: goldberg@parc.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Glycans are carbohydrates that are attached to proteins and lipids. They are constructed from simple sugars called monosaccharides with the sugars linked to form tree topologies. Just as the protein composition of a cell varies from tissue to tissue, so does the distribution of glycans. But unlike proteins, which are encoded in DNA, glycans are not constructed from a template. Instead the glycans present in a cell depend on which glycan-constructing enzymes (glycosyltransferases) are active in that cell. Glycans coat most cell surfaces in higher animals, and they play a crucial role in cell-to-cell signaling, including such important processes as sperm–egg binding and immune system recognition. Cancer cells produce aberrant glycans (Kobata and Amano, 2005), and several studies have suggested that glycans may serve as plasma biomarkers for cancer (An et al., 2005; Kirmiz et al., 2007).

Glycans can be identified at several levels of detail. The basic composition level specifies the number and types of monosaccharides, for example, five hexoses, five N-acetyl hexosamines (‘HexNAcs’) and two deoxyhexoses (fucoses). A more precise composition level specifies the types of hexoses (such as mannose, galactose or glucose) and HexNAcs (such as N-acetyl galactosamine or N-acetyl glucosamine). The ‘cartoon’ level gives the precise composition along with the connectivity (tree topology) of the monosaccharides, but not the exact types of linkages (glycosidic bonds) between adjacent monosaccharides. This intermediate level is perhaps the most commonly used level, because further detail can currently only be obtained with low-throughput techniques such as repeated rounds of mass spectrometry (MS) or treatment with enzymes that specifically digest certain glycans.

A comparison of glycan and peptide identification exposes some of the bioinformatics issues:

- Glycan databases lag far behind protein databases, because protein databases can be derived from DNA sequences.
- The ‘alphabet’ for most mammalian glycans is smaller than the alphabet for peptides, with only three to seven monosaccharides rather than 20 amino acids. Indeed the basic composition of a glycan can often be inferred from its total mass. The analogous task for peptides is currently impossible, because a peptide mass typically matches hundreds of distinct amino acid compositions within the resolution of MS instruments.
- Glycans are tree structures, and there are numerous isomeric forms—topologies—for any given composition. N-glycans share a common core, and their variable parts appear to follow a constrained set of patterns. Peptides are linear structures, and any sequence is possible.

Tandem-MS greatly outperforms single-MS for peptide identification, because a fragmentation spectrum contains much more information than a single-MS peak, which cannot even
The only program known to the authors for single-MS identification of glycans is Cartoonist, described by Goldberg et al. (2005). This program offers accurate, automatic annotation and hence the Consortium for Functional Glycomics (CFG, 2008) is now using it for glycan ‘profiling’ of various organisms and tissues by MALDI-TOF spectrometry. Cartoonist process single-MS glycan spectra in three steps.

1. The first step finds peaks in the spectrum that match the masses and isotope ratios of glycans. (We refer to the isotope envelope of a glycan as a single peak, rather than as a peak series, to avoid confusion with biosynthetic series.)

2. The second step generates all biologically plausible cartoons for each peak, using constraints believed to hold for all samples in the experiment.

3. The third step applies expert-system knowledge to score the cartoons generated in the second step. It then annotates each peak with its most likely cartoon(s).

The first two steps employ a ‘fixed’ table of monosaccharides and generic biosynthetic rules. The third step uses a ‘variable’ table input by an expert user; these rules encode information particular to the species, tissue or other special conditions of each spectrum in the form of scoring ‘demerits’. For example, Gal-Gal is not expected to occur in humans, and NeuGc is excluded from the brains of most mammals.

The field of glycomics has gathered expert knowledge over a number of years, yet still very little is known about most species and tissues; thus there is a need for a tool that avoids reliance on this knowledge. The goal of this article is to replace or augment demerits with a global analysis of a single-MS spectrum that exploits the connections between peaks. (Cartoonist currently annotates peaks individually, and makes no inferences from ‘ladders’ of peaks differing by the masses of mono- or disaccharides.)

In Figure 1, the glycan with monoisotopic mass 2111.06 Daltons differs from the glycan with mass 2285.15 by the addition of a single monosaccharide (the core fucose represented by the grey triangle). The glycan at 2489.25 Da can be similarly obtained from the one at 2285.15. Notice, however, that not all the glycans in this spectrum can be linked into a single ladder, and the similarities among the glycans are quite complex. By incorporating global, glycan family analysis, we aim to extend single-MS annotation to non-experts, and to samples for which little prior knowledge exists. Single-MS annotations are of necessity only preliminary, but can be extremely useful in quickly giving a sense of a sample, and in serving as a guide to more definitive analyses.

We developed three different family analysis algorithms, and tested our algorithms on 71 manually annotated spectra, publicly available on the CFG website (CFG Profiling data, 2008). The manual annotations are primarily based on single MS, but also use tandem-MS and painstakingly gathered human knowledge of the samples. (The manual annotations have also been checked, and
in a few cases corrected, by Cartoonist.) We refer to the manual annotations as the ‘ground truth’, even though some of them may be incorrect. In particular, most of the peaks are annotated with a single cartoon, even though additional isobars may be present.

2 METHODS

In this section, we first give the necessary background knowledge on mammalian glycans, essentially the information contained in the fixed table used to generate all biosynthetically plausible cartoons. We then describe three different algorithms: Parsimony, Random Walk and Max Subgraph.

2.1 Encoding mammalian N-glycans

A mammalian glycan is usually constructed from seven different types of monosaccharides (simple sugars), which we abbreviate as shown in Table 1. Most common glycan modifications of proteins fall into one of two broad categories: N-glycans, which are linked to the nitrogen in the side chain of asparagine, and O-glycans, which are linked to an oxygen in serine or threonine (Brooks et al., 2002; Taylor and Drickamer, 2006). In this article, we consider only N-glycans, although the techniques should also apply to O-glycans. N-glycans share a common core, containing two GlcNAc’s and three mannoses. N-glycans contain up to four (or rarely five) antennae growing from the mannoses, as illustrated subsequently.

We write antennae from bottom to top, and encode an N-glycan topology as a slash-separated list of antennae. Thus the left N-glycan as illustrated is encoded as ng/ngs/ngs/nt. By convention, if a glycan has only a single antenna, it is assigned to the second slot. Two antennae are assigned to the second and third slots, and three antennae to the first, second and third slots. Thus the cartoon at mass 2070.04 Da in Figure 1 is /ng/ngs/nt. This simple notation minimizes the ‘boilerplate’ present in other nomenclatures, but is only applicable to commonly occurring glycans.

Antennae are usually linear but sometimes have fucose (f) or GalNAc (o) attached. These ‘pendant’ monosaccharides are coded after the monosaccharide to which they are attached, so that the right-hand glycan above is encoded by ng/ngs/ngs/ntf.

The complete rules for generating valid (biosynthetically plausible) N-glycan antennae form a graph grammar (Fig. 2). Any of the following ‘terminal strings’ are valid: \( \epsilon \) (the empty string), n, ng, nfg, nfgf, nfgs, ngs, ngg, no, nfo, ngos and nos. In addition, any valid antenna \( \sigma \) can be made into longer valid antennae \( ng\sigma \) and \( nfg\sigma \) by preceding it with ng or nfg. Finally, any s in a valid antenna can be replaced by t. The terminal strings are based on the biology of mammalian glycans, and can be changed via a configuration file.

There are also two special cases, which we treat as optional one-letter antennae: bisecting GlcNAc (a monosaccharide tucked into the fork of the core), and core fucose (a pendant monosaccharide at the core GlcNAc). We encode bisecting GlcNAc and core fucose by b and f, respectively, after the fourth slash. Thus in Figure 1, the cartoon at mass 1835.92 Da is /n/n//f. The codes for the top row of cartoons: /ng/ngs/ntf. The one at mass 2111.06 is /ng/n//b. The one at mass 2285.15 is /ng/n//bf.

What we have described up until this point are the so-called complex N-glycans. There are two other classes of N-glycans as well. High-mannose glycans contain antennae composed of a single type of monosaccharide—mannose. In Figure 1, the peaks at 1987.98, 2192.08 and 2396.18 Da are high-mannose N-glycans.

Hybrid N-glycans contain one high-mannose branch and one or more complex antennae. In Figure 1, the peaks at 2029.01 and 2652.32 Da are hybrid N-glycans. We denote the right-hand branch in these two cartoons by H2, meaning hybrid with two mannoses. The cartoon code for the peak at 2029.01 Da is /ng/H2//. Conventionally, we place the high-mannose branch in the third antenna slot and leave the fourth antenna slot empty. See Figure 3 for more cartoon codes.

Table 1. The seven common monosaccharides in mammalian glycans, with letter codes and graphic symbols (NeuAc and NeuGc are both sialic acids)

<table>
<thead>
<tr>
<th>Code</th>
<th>Abbreviation</th>
<th>Full name</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>Man</td>
<td>Mannose</td>
</tr>
<tr>
<td>n</td>
<td>GlcNAc</td>
<td>N-acetyl-D-glucosamine</td>
</tr>
<tr>
<td>g</td>
<td>Gal</td>
<td>Galactose</td>
</tr>
<tr>
<td>s</td>
<td>NeuAc</td>
<td>N-acetyl-neuraminic acid</td>
</tr>
<tr>
<td>t</td>
<td>NeuGc</td>
<td>N-glycolyl-neuraminic acid</td>
</tr>
<tr>
<td>f</td>
<td>Fuc</td>
<td>Fucose</td>
</tr>
<tr>
<td>o</td>
<td>GalNAc</td>
<td>N-acetyl-D-galactosamine</td>
</tr>
</tbody>
</table>
In this project, we considered two cartoons to be equivalent if they contained exactly the same antennae (including bisecting GlcNAc and core fucose) in any order. Thus we considered ng/ngs/ng/ to be equivalent to ng/ngs/ngs/; even though in the first cartoon the antenna containing NeuAc is on a mannose with two filled antenna slots, and in the second cartoon, it is on a mannose with only one filled antenna slot. Distinguishing such similar glycans is very difficult with single-MS, and even difficult for tandem-MS, so we treat differences in antenna order like linkage information, and do not attempt to distinguish them.

2.2 The sample spectra
We tested our algorithms on manually annotated spectra, publicly available on the CFG website (CFG Profiling data, 2008). There were some groups of very similar spectra, and after selecting only one from each such group we obtained a test set of 71 samples (see the Supplementary Material for a full list). The 71 spectra range from 7 to 96 peaks, with a mean of 40. We do not consider ‘trivial’ peaks that have only one cartoon specified by the graph grammar. This includes high-mannose peaks, which although they can have multiple structural isomers are represented by a single cartoon. The peaks thus removed were primarily high-mannose peaks such as 1987.98, 2192.08 and 2396.18 Da in Figure 1. In that spectrum, the only other peak with a unique valid cartoon is the one at 2852.40.

The set of all 71 spectra contained a total of 2855 peaks. Of these 16% had only one biosynthetically plausible cartoon generated by the grammar. For the remaining peaks, the median number of annotations per peak was 21, with a max of 594.

2.3 Parsimony
Our first algorithm, Parsimony, exploits the biological fact that glycans are constructed by very specific enzymes (glycosyltransferases), only a subset of which are typically active in any one tissue. For each peak, the Parsimony algorithm first constructs the list of all biosynthetically plausible cartoons with the right mass. These are the cartoons containing the standard N-glycan core, along with four valid (possibly empty) antennae and optionally including bisecting GlcNAc and core fucose. For the mouse thymus example, there are four possible cartoons for the peak at 1835.92 Da ([/n/nf/], /n/uf/, /nuf/ and /nuf/h), two for the peak at 2029.01 (/ng/H2// and /ng/H2/) and 15 for the peak at 2244.12 (/ng/uf//, /ng/uf/, /ng/uf/ and /ng/uf//), etc.). Figure 3 shows the possible cartoons for the peak at 2070.04 Da. For the 14 ambiguous peaks, there are a total of 200 possible cartoons, containing 29 different antennae.

Next, Parsimony exploits the fact that the cartoons are not arbitrary trees, but contain a common core and a series of essentially linear antennae. As a rough approximation, different antennae are synthesized using different enzymes, so this algorithm finds a parsimonious set of enzymes by finding a smallest set of antennae that can be used to construct at least one cartoon for each peak. With unfixed numbers of peaks and antennae, this problem is NP-complete (transformation from vertex cover), but in practice this problem can be solved by exhaustive search. We first compute each required antenna (if any), meaning an antenna that appears in every cartoon for some peak. Then we try all possible sets of optional antennae, all sets of one, all sets of two and so forth, until we find all the minimal sets of antennae (required plus optional) that can construct one cartoon for each peak.

Finally, Parsimony annotates each peak with the cartoon(s) that can be constructed from some minimal set of antennae. A minimal set of antennae roughly corresponds to a minimal set of biosynthetic enzymes (glycosyltransferases), and hence a most parsimonious explanation for the observed peaks. This correspondence is not exact, because there is not one enzyme for each type of antenna; a typical glycosyltransferase recognizes the geometry of the last monosaccharide in an antenna and attaches one more monosaccharide, with a particular linkage.

| Table 2. Output of the Parsimony algorithm for the thymus example a |
|-----------------|-----------------|-----------------|-----------------|
| 1835.92 | /n/nf/ | |
| 2029.01 | /ng/H2// | |
| 2070.03 | /ng// | n/H2// |
| 2111.06 | ng/n// | |
| 2244.12 | /ng//f | n/H2//f |
| 2285.15 | ng/n//f | |
| 2489.25 | ng/n//f | |
| 2605.30 | /ng// | nfgs// |
| 2635.31 | /ng//f | n/fgf// |
| 2652.32 | ng/H2//f | |
| 2693.35 | ng/ng//f | |
| 2699.36 | ng/ngs//f | |
| 2839.41 | /ngfgs//f | nfgs// |
| 2932.40 | /ngfl| |
| 3026.49 | /ngfl//f | flfgfl// |

aIn the case of a tie, the algorithm output multiple predictions. Predictions matching ground truth are shown in bold.

2.4 Cartoon graphs
The other two algorithms we tested, called Random Walk and Max Subgraph, also start by assigning to each peak of the spectrum a list of all biosynthetically plausible cartoons of matching mass. These algorithms then use this list as the vertex set for a graph that we call a cartoon graph.

Glycans are constructed by the addition of monosaccharides, with little or no removal of monosaccharides in the construction process (other than in the initial synthesis of the trimannosyl core). Thus samples will usually contain not only a high-mass ‘mature’ glycan, but also many of the smaller glycans that are intermediate in the biosynthesis. So a correctly annotated single-MS spectrum is likely to contain ladders or even a graph of cartoons, in which each successive parent–child pair of cartoons differs by a single monosaccharide.

The simplest possible cartoon graph would place an edge between a pair of cartoons if the heavier cartoon differs from the lighter by the addition of a single monosaccharide, which could be an addition to an antenna, an addition of a core fucose or a bisecting GlcNAc or an addition of a pendant antenna fucose or GalNAc. Some enzymatic reactions, however, are much faster than others, so it is relatively rare to see an uncapped GlcNAc (that is, an antenna ending in n), and much more common to see each n in an antenna followed by a g. Thus, a better cartoon graph also places an edge between a pair of cartoons if the heavier differs from the lighter by the addition of the common disaccharide ng (lactosamine) to some antenna. We refer to this better cartoon graph as the 1-hop cartoon graph. One can carry this reasoning still further, and place an edge between two cartoons if the heavier differs from the lighter by the addition of any two monosaccharides. This is the 2-hop cartoon graph. The largest cartoon graph includes an edge between two cartoons if the heavier contains the lighter as a proper subgraph. This graph is the transitive closure cartoon graph.

As explained below, we also weighted edges within cartoon graphs, so that a ‘long’ edge, meaning one that required the addition of several
monosaccharides, had lower weight than one requiring only a single monosaccharide.

2.5 Max Subgraph

Max Subgraph favors cartoons that can be linked into a richly connected biosynthetic family. In order to match each peak to only one glycan, we seek a subgraph of the cartoon graph that contains exactly one vertex for each peak. So if the spectrum has $n$ peaks, Max Subgraph seeks the set of $n$ vertices (cartoons), one for each peak, that induces a subgraph with largest total edge weight.

This maximum-edge-weight problem is NP-complete, and this time the brute-force solution is too slow to be used, so we investigated several heuristics. The most successful heuristic consists of iterations of the following process. We pick an initial vertex at random for each peak of the spectrum, giving an initial subgraph. Then we hill-climb: for each peak (considered in random order) we systematically score the subgraphs with all possible vertices for that peak, and we choose the subgraph with the maximum total edge weight. There are often several maximum-weight subgraphs. In this case, we pick one at random; this detail turns out to be important. We continue to hill-climb until no further progress can be made. The process just described is then iterated, typically 100 times. At the end we have a set of subgraphs tied for the maximum edge weight. These subgraphs are not guaranteed to be optimal, but for simplicity we refer to them as maximal subgraphs.

Paralleling the Parsimony algorithm, the final annotations for a peak will be all the cartoons that appear in any maximal subgraph. Hence we attempt to compute all the maximum-weight subgraphs, not just a single maximum-weight subgraph. To do this, we systematically examine all perturbations of the maximal graphs found during the hill-climbing in order to find additional maximal subgraphs.

The algorithm was tested with several different versions of the cartoon graph and several different weighting schemes for the edges of the cartoon graph. The scheme that worked best was the transitive closure cartoon graph with inverse length weighting. That is, an edge that links a cartoon with $k_1$ sugars to one with $k_2$ sugars is weighted by $1/(1 + |k_1 - k_2|)$.

Table 3 illustrates Max Subgraph on the mouse thymus tissue, comparing it to the ground truth. The Max Subgraph solution actually has a higher edge weight than the manual annotation. In the Max Subgraph solution, the peak at 2693.35 Da has changed from ng/ng/ng/ to ng/ng/g. This new cartoon

Table 3. Results of the Max Subgraph algorithm on the thymus example\(^a\)

<table>
<thead>
<tr>
<th>Mass</th>
<th>Cartoon</th>
<th>Weight</th>
<th>Cartoon</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1835.92</td>
<td>/nu/lf</td>
<td>2.22</td>
<td>/nu/lf</td>
<td>2.22</td>
</tr>
<tr>
<td>2029.01</td>
<td>/ng/H2/</td>
<td>0.25</td>
<td>/ng/H2/</td>
<td>0.25</td>
</tr>
<tr>
<td>2070.03</td>
<td>/ng/fg/</td>
<td>2.83</td>
<td>/ng/fg/</td>
<td>2.83</td>
</tr>
<tr>
<td>2111.06</td>
<td>/ng/gf/</td>
<td>0.83</td>
<td>/ng/gf/</td>
<td>1.08</td>
</tr>
<tr>
<td>2244.12</td>
<td>/ng/dg/</td>
<td>2.83</td>
<td>/ng/dg/</td>
<td>2.83</td>
</tr>
<tr>
<td>2285.15</td>
<td>/ng/dg/</td>
<td>0.50</td>
<td>/ng/dg/</td>
<td>0.83</td>
</tr>
<tr>
<td>2489.25</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.50</td>
</tr>
<tr>
<td>2605.30</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.00</td>
</tr>
<tr>
<td>2635.31</td>
<td>/ng/dg/</td>
<td>1.00</td>
<td>/ng/dg/</td>
<td>1.00</td>
</tr>
<tr>
<td>2652.32</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.00</td>
</tr>
<tr>
<td>2693.35</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.00</td>
</tr>
<tr>
<td>2839.41</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.00</td>
</tr>
<tr>
<td>3026.49</td>
<td>/ng/dg/</td>
<td>0.00</td>
<td>/ng/dg/</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\(^a\)Differences between manual and automatic annotations are shown in bold. The Weight column is the sum of the edge weights extending from that cartoon to larger cartoons.

2.6 Random Walk

The Random Walk algorithm is motivated by viewing the synthesis of glycans as a Markov chain. When a glycan is constructed there are typically multiple options, that is, several different monosaccharides that could be added next. The Markov chain picks one of these options uniformly at random. In order that the Markov chain be strongly connected, we consider edges to be undirected, so that a heavier glycan can transition to a lighter glycan, as well as the other way around. The stationary distribution of the Markov chain can then be loosely interpreted as the ‘probability’ of each cartoon.

Table 4 shows the results of the Random Walk algorithm on the thymus example. Unlike Parsimony and Max Subgraph, Random Walk ranks its cartoon annotations, best, second best and so forth. In this example, however, the ground truth (marked in bold) was the first choice only 4 out of 11 times. There are 11 rather than 14 peaks in this case because the 1-hop cartoon graph is actually disconnected. Three of the peaks had no cartoons that directly linked (via a single monosaccharide of the disaccharide ng) to a cartoon on another peak.

To find the stationary distribution of the Markov chain, we build a sparse $n \times n$ matrix $M$ of transition probabilities, where $n$ is the number of biosynthetically plausible cartoons for all the peaks in the spectrum, which is typically around 1000. The entries of $M$ correspond to the (undirected) edges in the cartoon graph. We assume an initial uniform distribution $v$, and then compute $M^tv$ until convergence (about 50 iterations). $M^tv$ gives the stationary distribution, that is the probability distribution ‘at infinity’ of the location of a random walker.

As in the case of Max Subgraph, we tested the Random Walk algorithm on several different versions of the cartoon graph and with several different edge-weighting schemes. The 2-hop cartoon graph slightly outperformed the other versions of the cartoon graph. Disconnected 2-hop cartoon graphs were rare, but in the case of an isolated peak with no connections, the stationary distribution was simply the initial uniform distribution. Edge weighting had little effect with one exception. We found that performance improved when a correction was added for multiedges. For example, there are three ways to obtain the cartoon ngng/ng/ from ng/ng/ng/, corresponding to the three antennas that could receive the addition of ng, so the multiedge correction gives this edge a weight of 3. Without the multiedge correction, Random Walk gave top choices with excessive antennae heterogeneity.

Up to this point, we have ignored the possibility that a sample might contain multiple glycans of the same mass (isobars). At the very least an

Table 4. Results of the Random Walk algorithm on the 1-hop graph using the thymus data\(^a\)

<table>
<thead>
<tr>
<th>Mass</th>
<th>/no/1/</th>
<th>/no/1/</th>
<th>/no/1/</th>
<th>/no/1/</th>
<th>/no/1/</th>
</tr>
</thead>
<tbody>
<tr>
<td>1835.92</td>
<td>/no/1/</td>
<td>/no/1/</td>
<td>/no/1/</td>
<td>/no/1/</td>
<td>/no/1/</td>
</tr>
<tr>
<td>2070.03</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2111.06</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2244.12</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2285.15</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2489.25</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2605.30</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2635.31</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2652.32</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2693.35</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>2839.41</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
<tr>
<td>3026.49</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
<td>/ng/1/</td>
</tr>
</tbody>
</table>

\(^a\)Each row gives the mass and top four states (cartoons), with the most probable state on the left. The ground truth cartoon is in bold.
3 RESULTS

Comparing the three algorithms to the ground truth and to each other is not completely straightforward. The manual annotations usually specify a single cartoon for each peak, although there are a few peaks with multiple ground truth annotations. On the other hand, Max Subgraph and Parsimony often give multiple annotations per peak, and Random Walk gives a ranked list of annotations per peak.

If we simply compare the number of times each algorithm’s set of annotations includes the correct cartoon, we obtain the numbers shown in Table 5. These results exclude all ‘trivial’ peaks, such as high-mannose peaks, for which the grammar generates only a single cartoon. If we included the approximately 400 trivial peaks, we would of course obtain higher percentages. Out of the 2396 non-trivial peaks in the 71 spectra, the average number of possible cartoons per peak is 53, and the median number is 21. Based on the median, a program that randomly guessed one cartoon for each peak would get a hit only 5% of the time. Thus all three algorithms performed quite well.

The simple comparison above, however, is unfair, because Parsimony generates more than three times as many cartoons per peak as Maxgraph, and thus has a better chance of hitting the correct cartoon. So to compare Max Subgraph with Parsimony in more detail, we use a performance metric we call ability, which takes into account the number of proposed cartoons for each peak. Ability measures how much better the algorithm does than chance. For example, if there are 100 biosynthetically plausible cartoons and the algorithm proposed five cartoons, one of which agreed with the unique ground truth cartoon, we would say that the algorithm performed about 20 times better than chance. Suppose there are \( n \) biosynthetically plausible cartoons for a peak, the algorithm proposes \( k \) cartoons, and there are \( g \) ground truth cartoons. If the algorithm were guessing at random, the probability that none of the \( k \) cartoons includes even a single ground-truth cartoon is

\[
\frac{\binom{n-k}{k}}{\binom{n}{k}} = q.
\]

Then \( p = 1 - q \) is the probability of hitting a ground-truth cartoon for this peak by chance. Now in order to integrate over all the peaks in the spectrum, we count the fraction of peaks for which the algorithm found a ground-truth cartoon, and we divide this fraction by the median of the \( p \) values for the peaks on which the algorithm found a ground-truth cartoon.

The ability scores of Parsimony and Max Subgraph on our 71 test spectra are shown in Table 6. Max Subgraph outperforms Parsimony in both mean and median. Max Subgraph usually returned a single cartoon for each peak, whereas Parsimony typically returned two.

We then performed a more detailed comparison of Max Subgraph with Random Walk. Because Random Walk ranks its answers from ‘most probable’ on down (with some number of ties), this comparison is relatively straightforward. On a peak for which Max Subgraph returns \( k \) cartoons, we examined the top \( k \) cartoons found by Random Walk, and declared that Random Walk was correct if one of these \( k \) agreed with a ground-truth cartoon. The result was that Max Subgraph outperformed Random Walk about two-thirds of the time. On 44 of the spectra, Max Subgraph outperformed Random Walk, while on 22 of the spectra Random Walk did better; and on five of the spectra the two algorithms were tied. Random Walk also tended to fail more catastrophically in the cases that it failed.

4 CONCLUSIONS

Our investigations show that using general biological principles of glycan synthesis can make the automatic interpretation of single-MS spectra of glycans quite practical. We investigated three different ways of encoding these principles into a peak annotation algorithm, and each performed significantly better than random guessing. Indeed each algorithm performed at least five times better, as measured by the chance of including a ground truth cartoon among the set of annotations for a randomly picked peak. Max Subgraph using the transitive–closure cartoon graph was judged to be the overall best of the algorithms we tested. One more way of looking at the performance: Max Subgraph uniquely assigns the correct cartoon to more than half of the peaks in 41 out of the 71 spectra.

Glycan family analysis improves on original Cartoonist both in performance and usability. Family analysis gives a reasonable initial cartoon assignment without any sample-specific information. As argued above, we found a roughly 5-fold reduction in the number of ambiguous peaks. Perhaps even more important than this performance improvement is the improvement in usability. With glycan family analysis, the user directly specifies additional information, obtained from either expert knowledge or further experiments, by correcting initial cartoon assignments. The original Cartoonist requires corrections to be abstracted into the form of demerits. Cartoonist currently has 23 features (for example, presence of a bisecting GlcNAc) to which demerits may be applied, and more
such features will need to be added as it is applied to a wider range of organisms.

A peak given multiple cartoon annotations by family analysis is a candidate for further experimental analysis, which in turn can improve the annotation of other peaks. We performed a small computational experiment to explore this. We picked two peaks from each spectrum and ‘pinned’ them to their correct answer—i.e. simulated doing extra experiments (e.g. tandem-MS) to determine their cartoons. We ran an exhaustive search on 57 spectra, pinning all pairs of peaks. The number of correctly annotated peaks in Max Subgraph went from 1298 to 1601 out of 2047 peaks. So the percentage correct went from 63% to 77%, where the 77% does not include the pinned peaks. We believe that the algorithms we have presented make it practical to automatically construct rough-draft annotations from single-MS, and increase the effectiveness of any extra structural information that is available.

Conflict of Interest: none declared.

REFERENCES


Lapadula,A.J. et al. (2005) Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS