Exploring the Landscape of Glycan Structure

MSc Bioinformatics and Theoretical Systems Biology Project

Steven Hargreaves

CID 01453122

Imperial College London
Department of Life Sciences
Word Count: 5479

Abstract

Glycans are highly prevalent in living organisms, and play keys roles in a variety of important biological processes such as cell signalling, immune reaction, viral infections and tumour cell proliferation. The difficulty of analysing glycan structures experimentally provides the motivation to develop reliable computational analysis techniques. A hindrance to this effort though is the complex tree-like structure of glycans, making them more difficult to analyse and compare than, for example, the linear chains of polypeptides in proteins. We describe a novel technique for measuring glycan similarity based upon a consideration of the biological reactions necessary for their synthesis, and which avoids the need to model tree structures. By analysing the similarities between 4111 glycans from a public database, we demonstrate that we can identify communities of glycans sharing the same or similar motifs, and that we can predict the presence of specific motifs in glycans which do not as yet have motif annotations.

Acknowledgements

Many thanks to John Pinney for his supervision and guidance during this project, and to Suhail Islam for technical assistance. The database and JSON file icons used in Figure 4 were designed by Smashicons from Flaticon.

Contents

Abbreviations

GT: Glycosyltransferase

JSON: JavaScript Object Notation

RDF: Resource Description Framework

SPARQL: A recursive acronym for SPARQL Protocol and RDF Query Language

UPGMA: Unweighted Pair Group Method with Arithmetic Mean

WURCS: Web3 Unique Representation of Carbohydrate Structures

XML: Extensible Markup Language

Glossary

Anomericity: The configuration, either axial (α) or equatorial (β), of the OH-group at carbon number one of a monosaccharide.

Archeopteryx: Phylogenetic tree visualisation software.

Binary adjacency matrix: A square matrix in which elements represent the adjacency of pairs of vertices in a graph.

Glycan: Compounds of monosaccharides linked glycosidically, which exist in free form or in covalent complexes with proteins or lipids .

Glycan motif: A commonly occurring glycan substructure appearing in multiple glycans.

Glycosyltransferase: An enzyme which catalyses glycosidic linkages.

GlyTouCan: Online database of glycan structures.

Jaccard similarity coefficient: A statistic for comparing the similarity of sample sets.

PhyloXML: An XML language for the storage of phylogenetic tree data.

1 Introduction

1.1 Motivation

Monosaccharides are carbohydrates which cannot be further hydrolysed to simpler compounds, such as glucose, fructose, and galactose. Glycans, synonymous with polysaccharides, are compounds of monosaccharides (usually more than ten) linked glycosidically (McNaught & McNaught 1997), which exist in free form or in covalent complexes with proteins or lipids (Yamanishi et al. 2007). Evidence has been found relating glycans to numerous important biological functions, including immune reaction, protein stabilization, cell signalling (Bucior & Burger 2004), and embryogenesis (Rosa & Reinhold 2002). Several aspects of tumour progression, such as metastasis and angiogenesis, are regulated by glycans, and tumour cell proliferation is potentiated by glycans activating growth-factor receptor tyrosine kinases (Fuster & Esko 2005). Some pathogenic bacteria and viruses have been shown to infect their hosts via glycan-receptor interactions (Cossart & Sansonetti 2004, Sacks & Kamhawi 2001), whilst other research has demonstrated that glycan profiles can represent important signatures of disease states (Tong et al. 2003). Studies have also implicated glycans in the regulation of inflammatory response. For example, selectins and Sialyl Lewis^{X} oligosaccharides, two types of glycan, have been used to probe cell adhesion molecules, which when blocked can inhibit inflammation (Albelda et al. 1994).

Given the important and diverse functions attributed to glycans, the ability not just to determine their individual structure, but also to accurately quantify their presence or absence within different cell and tissue types, is of great value. However, for reasons we expand upon below, glycan structure analysis presents

major challenges. In this research we describe a novel computational approach allowing us to explore glycan similarity, by means of characterising glycans in terms of the glycosidic links between their constituent monosaccharides, rather than taking the more complex approach of modelling the complete glycan tree structure.

1.2 Glycan Structure

Glycans exhibit a tree-like structure, which can be described in terms of its component monosaccharides and the glycosidic links between them. Figure 1 shows an example glycan, in which the differently coloured and shaped nodes represent different monosaccharides, and the labelled edges represent the glycosidic links (see `https://www.ncbi.nlm.nih.gov/glycans/snfg.html` for a full list of monosaccharide symbols).

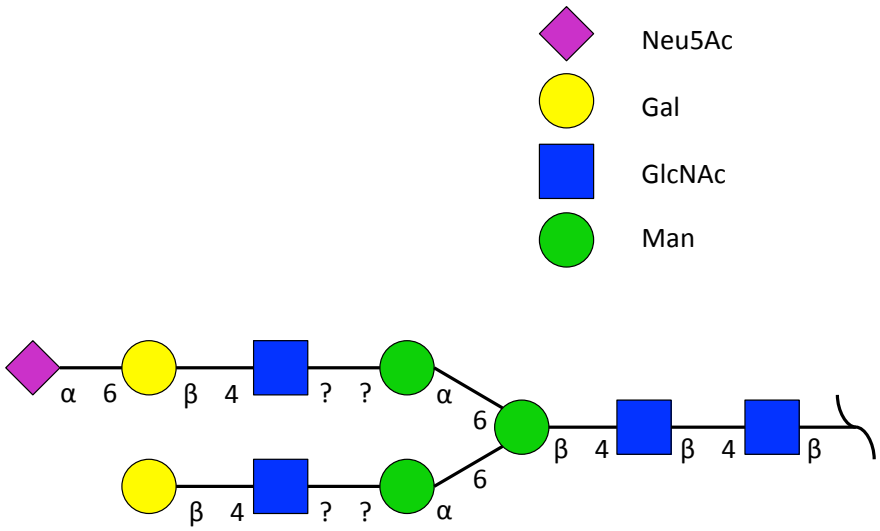


Figure 1: An example glycan diagram. The differently coloured and shaped nodes represent different monosaccharides, and the labelled edges represent the glycosidic links between them (image taken from the GlyTouCan repository[1]). Glycosidic links are characterised by the anomericity (α or β) of carbon 1 on the first monosaccharide, and the carbon number of the non-anomeric carbon on the second monosaccharide. In a number of published glycans, the nature of some glycosidic links has not been satisfactorily established, and is therefore denoted as ambiguous via the use of question mark characters.

When dissolved in an aqueous environment, the monosaccharides (or sugars) of which glycans are comprised, primarily exist in the cyclic (or hemiacetal) form. In this form, the sugars are further characterised by their anomeric configurations – either the α-anomer, with an axial OH-group at carbon 1, or the β-anomer, with an equatorial OH-group at carbon 1 (Song et al. 2012). The glycosidic links between the sugars exist between carbon 1 (the anomeric carbon) of one sugar, and some other, non-anomeric carbon of another. Hence the links are defined by their anomericity (α or β), and the carbon number of the non-anomeric

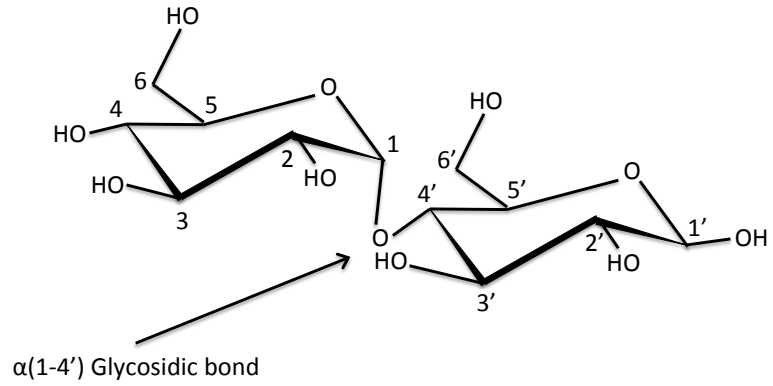
[1] `https://glytoucan.org/Structures/Glycans/G31576LD`

carbon on the second sugar. This can be seen more clearly in Figure 1, where, for example, the N-Acetyl-Neuraminic Acid (purple diamond) is glycosidically linked via its α-anomer carbon 1 to carbon 6 of the D-Galactose (yellow circle), and therefore labelled as 'α 6'. This anomericity is significant – two pairs of monosaccharides glycosidically linked via the same carbons but with different anomeric configurations result in two stereochemically distinct disaccharides, as illustrated in Figure 2. Here, the polymeric form of the maltose disaccharide (starch) is digestible by humans, whereas cellulose, formed from basic repeats of cellobiose, is not (Song et al. 2012).

1.3 Glycan Analysis

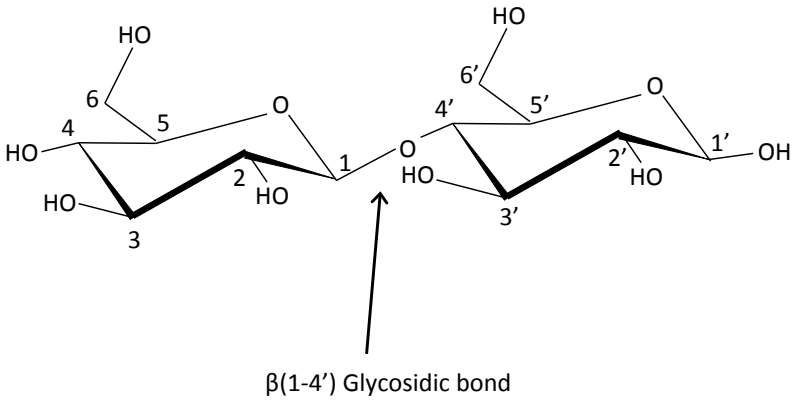
Methods of characterising glycan structures include mass spectral analysis (Bennun et al. 2013), high performance liquid chromatography, capillary electrophoresis, and nuclear magnetic resonance technology (Von Der Lieth et al. 2004), whilst gene expression processing may be used to characterise glycosylation processing enzymes (Bennun et al. 2013).

Glycan analysis though remains highly problematic – unlike the comparatively mature techniques regularly applied to DNA and RNA analysis, whose nucleotide chains are both linear and contain only 4 elementary components, and which are easy to amplify using (for example) polymerase chain reactions, glycans are tree-like in nature, and the large number of component monosaccharides involved together with the wide range of possible glycosidic links makes glycan structure determination challenging. Consequently only a small number of samples are available for analysis (Kawano et al. 2005).



[4-*O*-(α-D-Glucopyranosyl)-β-D-Glucopyranose]

(a) Maltose



[4-*O*-(β-D-Glucopyranosyl)-α-D-Glucopyranose]

(b) Cellobiose

Figure 2: Pairs of the same monosaccharide (Glucose) glycosidically linked with different anomericity form different disaccharides (a) maltose and (b) cellobiose. Image based on Figure 4 from (Song et al. 2012).

To mitigate for these experimental shortcomings, in much the same way as sequence homology and multiple alignments may be exploited to construct complete genomes from DNA fragments (Staden 1979), there is a need for reliable computational techniques of glycan structure and prevalence prediction. Examples of research towards this end include probabilistic modelling of glycan families (Ueda et al. 2005), identification of glycan fingerprint differences between prostate cancer cell stages (Bennun et al. 2013), prediction of glycan structures from gene expression profiles (Kawano et al. 2005), multiple glycan alignments by modelling glycans as trees (Hosoda et al. 2017), and glycan comparison (Aoki, Mamitsuka, Akutsu & Kanehisa 2004, Aoki, Yamaguchi, Ueda, Akutsu, Mamitsuka, Goto & Kanehisa 2004). A review of glycan data formats and toolboxes can also be found in (Campbell et al. 2014).

1.4 Glycan Diversity, Locations and Motifs

In contrast to the type of bonds found in proteins and nucleic acids, the high number of different glycosidic link configurations possible between monosaccharides leads to high variation and structural diversity between glycans, which confer distinctive characteristics to the cell surface where they are typically found (Bennun et al. 2013). In one online glycan database (GlyTouCan[2]), for example, there are a total of 105050 distinct glycan structures. Glycans do however share some common sub-structures, known as motifs. For example, the glycan structure represented in Figure 1 contains the 'N-Glycan core basic' motif, which is displayed in Figure 3. There are just 61 motifs presently listed in the GlyTouCan database - the full list

[2] `https://glytoucan.org/`

(with diagrams) is available via the footnote link[3].

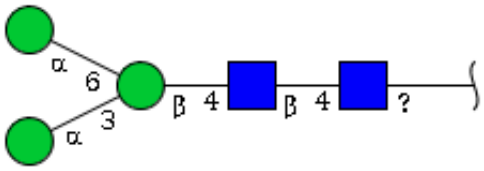


Figure 3: The 'N-Glycan core basic' glycan motif, which can also be seen as part of the structure of the glycan shown in Figure 1.

1.5 Proposed Method for Exploring the Landscape of Glycans

Intrinsic to the method of predicting of glycan structures from gene expression profiles cited above (Kawano et al. 2005), is a recognition of the fact that a specific set of glycosyltransferases (GTs) are required to catalyse the reactions necessary for synthesising specific glycans. By building a library of the bond formation patterns of GT reactions, and determining a co-occurrence score of the reaction patterns in a glycan database, the authors predicted glycan structures with an accuracy of 81%. Additionally, using gene expression data from the human carcinoma cell, the authors predicted the presence of sialic acid and the Sialyl Lewis X motif.

Expanding upon that approach, in this paper we describe a method of exploring glycan similarity from biochemical reaction similarity. More specifically, to compare any two glycans we calculate the Jaccard similarity coefficient between the two sets of biochemical reactions necessary for the synthesis of each glycan in the pair. We visualise clusters of similar glycans via the use of networks, which

[3] `https://glytoucan.org/Motifs/listAll`

allows us to explore whether or not glycans which we have judged to be similar in terms of their biochemical synthesis reactions are also similar in terms of their constituent motif sets. Furthermore, we perform hierarchical clustering using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), allowing us to form a dendogram, which can be explored interactively using bespoke phylogentic tree visualisation software and reveals large-scale groups of similar and dissimilar glycans.

The rest of this paper is organised as follows – in Section 2 we describe, in sufficient detail for a Bioinformatician or Computer Scientist to be able to reproduce our results, the databases and computational methods used, before presenting and discussing our results in Section 3. In Section 4 we discuss our findings within the bigger picture of glycan analysis, before providing suggestions for further work and our conclusions in Sections 5 and 6 respectively.

2 Method

Figure 4 illustrates, at a high level, the methods used in this research. In the following sub-sections we describe each of these steps in more detail.

2.1 Reproducibility

The computer program code referenced throughout this report, and which was used to produce the results described, is available from a GitHub software repository. Details of how to obtain the code, as well as the Python package dependencies, can be found in Appendix A.

2.2 Glycan Databases

A number of different publicly accessible glycan databases exist, as well as multiple different formats for expressing glycan structure. The details of these databases and formats are listed in Tables 4 and 5 in Appendix B. We use the GlyTouCan database, given its larger number of glycans and supported formats compared to the other databases, as well as the fact that it also contains motif information. In our trials it was also found to have the most suitable query method. We also chose to work with the Web3 Unique Representation of Carbohydrate Structures (WURCS) glycan structure format, which is the most recently devised format and which has been designed to address various shortcomings of the other existing formats (Matsubara et al. 2017).

GlyTouCan is a semantic web graph database, rather than a relational database. Such databases contain Resource Description Framework (RDF) data, in which all data items take the form of *triples*, and the database is queried using the query

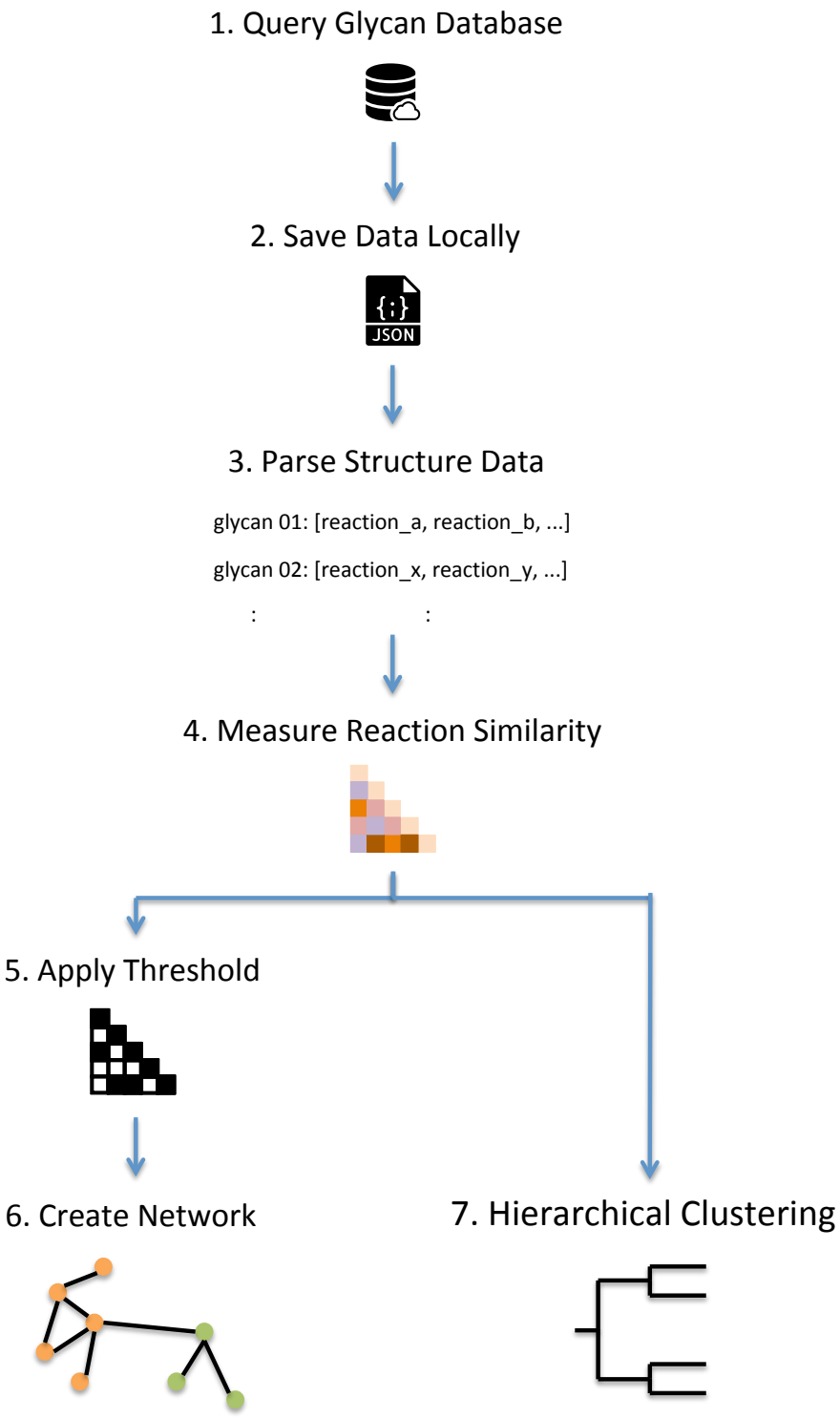


Figure 4: Method flowchart

language SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language). A full description of this type of data is beyond the scope of this report, however some useful primers are available (W3C Working Group 2014, W3C SPARQL Working Group 2013).

The data we wish to extract from GlyTouCan is a list of glycan IDs, the text string (in WURCS format) describing the structure of each glycan, and the IDs of the motifs which each glycan contains. Separately, we also wish to extract a list of the motif names associated with each motif ID. The two SPARQL queries we use for these tasks are given in Appendix C, and the Python scripts used to execute them are called `glytoucan_rdf_to_json.py` and `motif_rdf_to_json.py` respectively. After execution the results are saved locally to JavaScript Object Notation (JSON) files.

2.3 Extracting Reactions from Glycan Structure Data

As described in Section 1.5, we wish to characterise our glycans in terms of the biochemical reactions necessary for their synthesis. We do this by parsing the WURCS string describing each glycan structure, and, recalling that glycans exhibit a tree-like structure, we extract collections of 'child' and 'parent' monosaccharides, and the type of the link that joins them (i.e. the carbon numbers of the child and parent) for every glycan. The full WURCS format specification can be found in (Matsubara et al. 2017), and we provide an illustrative example in Figure 13 Appendix D.

For some of the glycan entries in the database there is uncertainty regarding aspects of the glycan structure – for example the number of links or the carbon

numbers expressed with only a percentage of certainty. We reject these. The Python script which parses the WURCS strings is `parse_glytoucan_results.py`.

2.4 Filtering the Glycan Dataset

The previous step produces 33142 glycans. As we shall see in the next step, we wish to create a similarity matrix for these glycans – that will be a 33142×33142 matrix of 64 bit floats, which requires approximately 8 gigabytes of RAM. This amount of RAM, whilst easily obtainable on current server clusters, is approaching the limit of desktop machines. Additionally, the computation time required for calculating all of the values in such a matrix is excessive. Consequently, using the Python script `filter_by_common_motifs.py`, we perform a filtering step on the dataset. We determine the top 10 most frequently observed motifs (from single-motif glycans), and eliminate from our our list any glycans whose motifs are contained within the top 10 list. The result is a more tractable list of 4111 glycans.

2.5 Calculating Distances Between Glycan Reaction Collections

We determine a measure of glycan similarity using two alternatives methods, which we compare in Section 3. Firstly, by calculating the Jaccard similarity coefficient between all pairs of reaction *sets* (that is, we only count each reaction belonging to a glycan once, regardless of whether or not it occurs multiple times), and secondly, by calculating a modified version of the Jaccard similarity coefficient between all pairs of reaction *lists* (i.e. in this case we do count the number of times each reaction occurs).

The Jaccard similarity coefficient between two sets A and B is given by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

As an example, consider two Glycans, G_1 and G_2. Let G_1 contain the set of reactions r_1, r_3, r_5 and r_5, whilst G_2 contains reactions r_4 and r_5 (neither glycan contains reaction r_2). We want to count the number of reactions shared by G_1 and G_2 as a fraction of the total number of unique reactions present in both sets. Table 1 illustrates the two sets.

Glycan	**Reaction**			
	r_1	r_3	r_4	r_5
G_1	1	1	0	2
G_2	0	0	1	1
Present in both	0	0	0	1

Table 1: Jaccard similarity coefficient calculation for glycans G_1 and G_2.

In this case the two glycans have just one reaction in common out of a total of four different reactions present in both glycans, and so the Jaccard similarity coefficient is 1/4, or 25%. In the modified Jaccard case, where we wish to take into account reaction quantities, instead of scoring one when both glycans share a reaction we calculate the score for the shared reaction as the smaller reaction quantity divided by the larger one, as illustrated in Table 2.

The modified Jaccard similarity coefficient is then $\frac{1/2}{4}$, or 12.5%. Hence a penalty is incurred when glycans share reactions, but in unequal quantities, with the penalty proportional to the quantity difference.

Glycan	Reaction			
	r_1	r_3	r_4	r_5
G_1	1	1	0	2
G_2	0	0	1	1
Shared fractional score	0	0	0	1/2

Table 2: Modified Jaccard similarity coefficient calculation for glycans G_1 and G_2.

For both methods, we calculate the similarity coefficient for all pairs of glycans, resulting in two similarity matrices. We will use these similarity matrices directly when performing hierarchical clustering as described in Section 2.7, but for the purposes of creating a network of glycans to explore, we separately create a binary adjacency matrix by applying a heuristically chosen threshold to the values in the similarity matrix, such that any values falling below the threshold now score 0% similarity, whilst any falling on or above the threshold score 100% similarity.

2.6 Glycan Network Creation

The binary adjacency matrix created in the previous step enables us to form a network, or undirected graph, representing glycan similarity, in which each of our glycans is a node and two nodes are joined by an edge if their glycans are marked as similar in the adjacency matrix. We use the Python package NetworkX (Hagberg et al. 2008) for this purpose.

In our Python script `glycan_reaction_distances_to_network.py` we create a graph using the `from_pandas_adjacency` function from the NetworkX package. For the purposes of visualising this graph, we also need to assign each node to a point in the two-dimensional Euclidean plane. NetworkX provides several node positioning algorithms for this purpose, each of which takes a different approach to

laying out the nodes in two-dimensional space according to various network metrics. The choice of any particular algorithm here does not altar the network topology, but can subjectively result in more or less informative graph visualisations. We choose the Fruchterman-Reingold force-directed algorithm (Fruchterman & Reingold 1991).

We would like some method of judging if the relationships present in our network do indeed correlate with glycan similarity. From our original data query (see Section 2.2), we know which motifs are present in each glycan structure (according to the annotations in the GlyTouCan database). In the Python script `display_network.py` we use this information to colour each node according to the set of motifs its glycan contains, and accordingly we would hope to see nodes of equal colour appearing in obvious clusters or communities in our graph.

In Section 2.5 we described two alternative methods of measuring similarity between glycans – the Jaccard similarity coefficient applied to a glycan's reaction set, and the modified version applied to its reaction list (i.e. including reaction quantities). We asses the relative merits of these two methods via the `attribute_assortativity_coefficient` function in NetworkX, after setting a 'motif' attribute for each node, whose value is the list of motifs which that glycan contains. The assortativity coefficient measures the similarity of connections in the graph with respect to a given attribute (Newman 2003). However, the two graphs resulting from the two methods will almost certainly contain an unequal number of edges. In order to make the comparison fair, when creating the graph for the second method, rather than using the adjacency matrix, we use the similarity matrix (i.e. before thresholding has been applied) and the NetworkX function `from_numpy_array`, applying the similarity coefficients as edge weights. We then

restrict our graph to only the X highest weight edges, where X is the number of edges found in the graph created using the first method, thus ensuring that both graphs have an equal number of edges.

Given the amount of glycans in our dataset, the resulting graph is large, and therefore can be uninformative when visualised as a whole. We are particularly interested in discovering groups of glycans deemed to be similar to each other, and so we use the `connected_components` function in NetworkX to obtain lists of nodes which are connected to each other. Ordering these lists according to the number of nodes they contain allows us to select and visualise only the largest groups of connected components found in our graph.

2.7 Hierarchical Clustering

When plotted on a computer, the graphs described above can be explored by zooming in to focus on particular groups of nodes/glycans, potentially revealing interesting relationships between (for example) glycans sharing the same or similar motifs. The method though is dependent upon choosing a threshold value at which to make the binary choice between similar or dissimilar for any two glycans. An alternative method of clustering, avoiding this limitation, is UPGMA (Gronau & Moran 2007). This algorithm represents the structure of a pairwise similarity matrix as a rooted dendogram. The algorithm proceeds in s step-wise manner, calculating the average of the distances between all elements in two clusters A and B. At the first step, each individual element (glycan, in our case) of the similarity matrix forms an individual cluster. The most similar clusters in any step are combined to form a new cluster, and the process is repeated until all clusters have

been paired. The Python script `upgma_glycans.py` creates one of these trees using the `DistanceTreeConstructor.upgma` function from the BioPython (Cock et al. 2009) package.

The resulting tree is given in the form of an XML file, specifically phyloXML (Han & Zmasek 2009). In order to aid visualisation, in our script `colour_tree_by_motif_group.py` we apply colour labels to all terminal nodes according to the motif sets belonging to the glycan represented by each node. We then traverse the tree from the terminal nodes upwards, applying the same colour labels if and only if the two children of a parent node have been labelled with the same colour. This provides a useful at-a-glance mechanism of identify clades of glycans sharing motif groups. Finally, we use the phylogenetic tree visualisation software Archeopteryx (Han & Zmasek 2009) to visualise and explore the tree.

3 Results

Possibly the most effective way to explore the landscape of glycan structure using the techniques described above in Section 2, is to manipulate them in real-time on a computer, where it is possible to zoom in to examine regions of interest and to (in the case of the Archeopteryx tree viewer) re-draw the tree in a multitude of different ways. In this section we present a selection of static images as we explore the glycan network and tree, illustrating both the validity of the techniques and some of the insights into glycan structure they provide.

3.1 Glycan Network

Figures 5 through to 9 below were all produced using the reaction set method (see Section 2.5). Starting from a total of 105050 glycans in the GlyTouCan database, this number is reduced to 33142 after rejecting glycans with structural uncertainties, and finally further reduced to 4111 glycans when we consider only glycans containing motifs from the list of top 10 most frequently occurring motifs (see Sections 2.3 and 2.3).

Figure 5 shows the most connected component of the network, with 444 nodes (representing glycans) coloured according to the sets of motifs they contain. The figure legend details the colours assigned to each motif set. It is immediately apparent that, as hoped for, same coloured nodes exist in distinct communities, with mutual connections and often in close proximity, indicating high numbers of connections between glycans sharing the same motif sets. The fact that individual nodes of the same colour are still spatially separate indicates that despite sharing the same motifs, there are nevertheless structural differences between those gly-

cans. This is exactly as we would expect, given that all glycans taken from the database are structurally distinct, albeit many show similarity in terms of their motifs. In order to examine the relationships between these groups more closely, we need to zoom in and inspect the network in more detail.

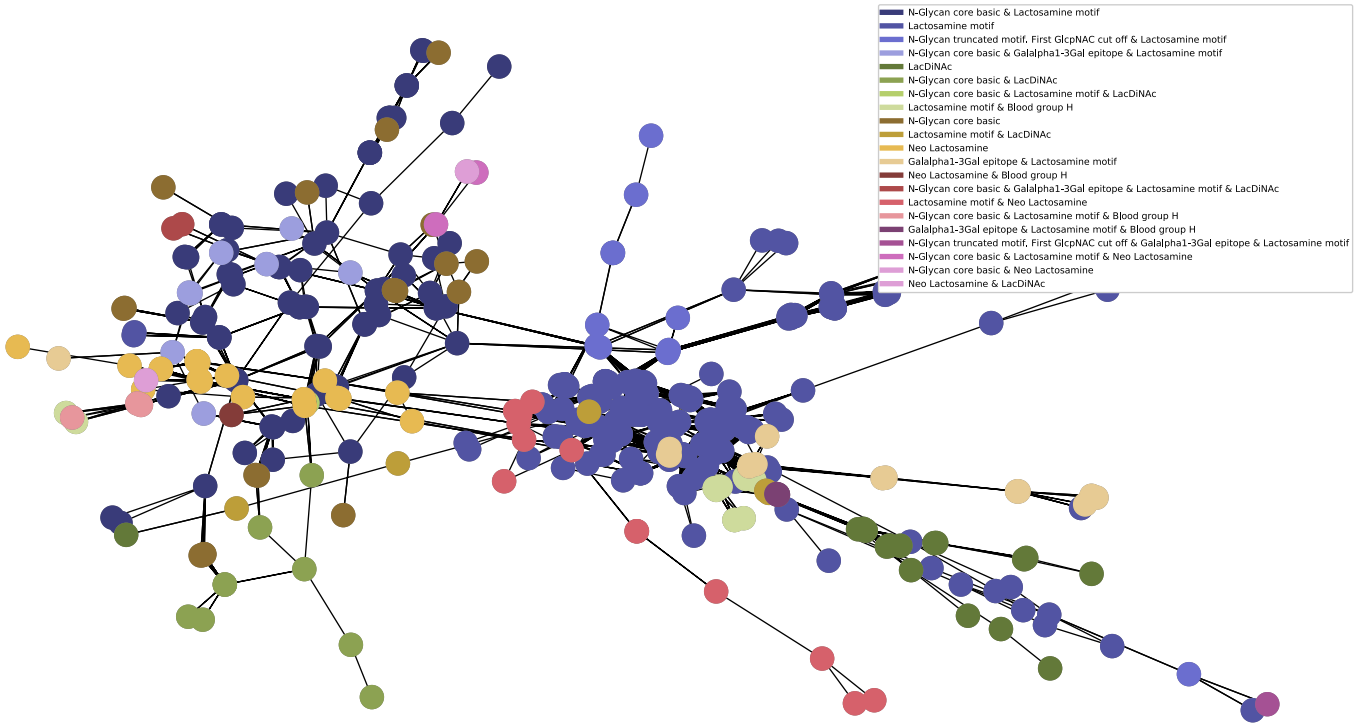


Figure 5: Most connected component of a network of glycans, with similarity determined using the reaction set method.

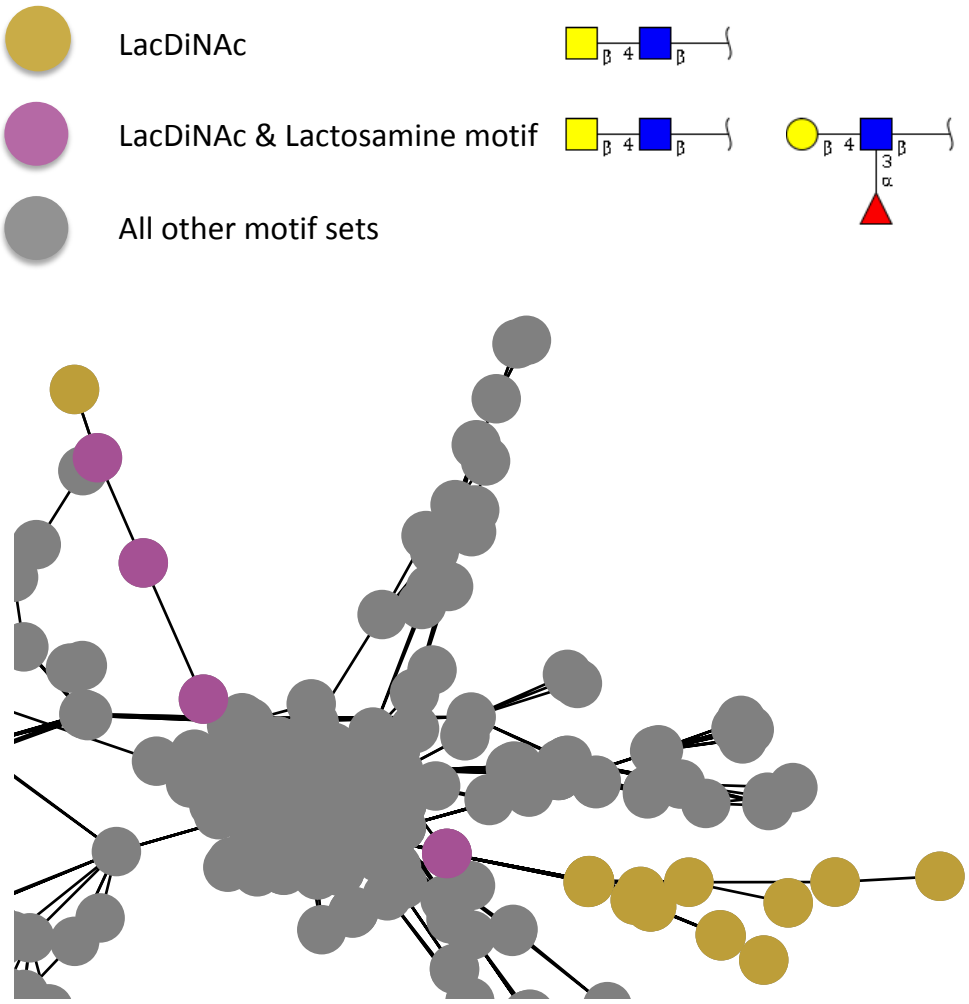


Figure 6: Glycans containing the LacDiNAc motif (orange nodes) are connected solely to glycans containing the LacDiNAc and Lactosamine motifs (purple nodes).

In Figure 6, we highlight only two groups of glycans – those containing the single motif LacDiNAc (coloured orange), and those containing two motifs – LacDiNAc and the Lactosamine motif (coloured purple). All other glycans are coloured grey. In the same figure, we also display the glycan motifs. The two motif sets are similar to each other in two respects – firstly, they both share exactly the same motif, LacDiNAc, and secondly, Lactosamine and LacDiNAc have in common a

β anomeric N-Acetyl-D-Glucosamine monosaccharide (the blue square). It is reasonable then to assume that glycans containing either of these motif sets would be structurally similar, and indeed the figure illustrates that the glycans in our network with just the LacDiNAc motif (orange) are solely connected either to each other or to glycans containing the LacDiNAc and Lactosamine motifs.

Another example is shown in Figure 7, this time highlighting glycans containing the Galalpha1-3Gal epitope, Lactosamine, and Blood group H motifs (orange), and those containing Lactosamine, and Blood group H motifs (green). Direct connections exist between the two groups, as we would expect given that one group contains two out of the three motifs present in the other.

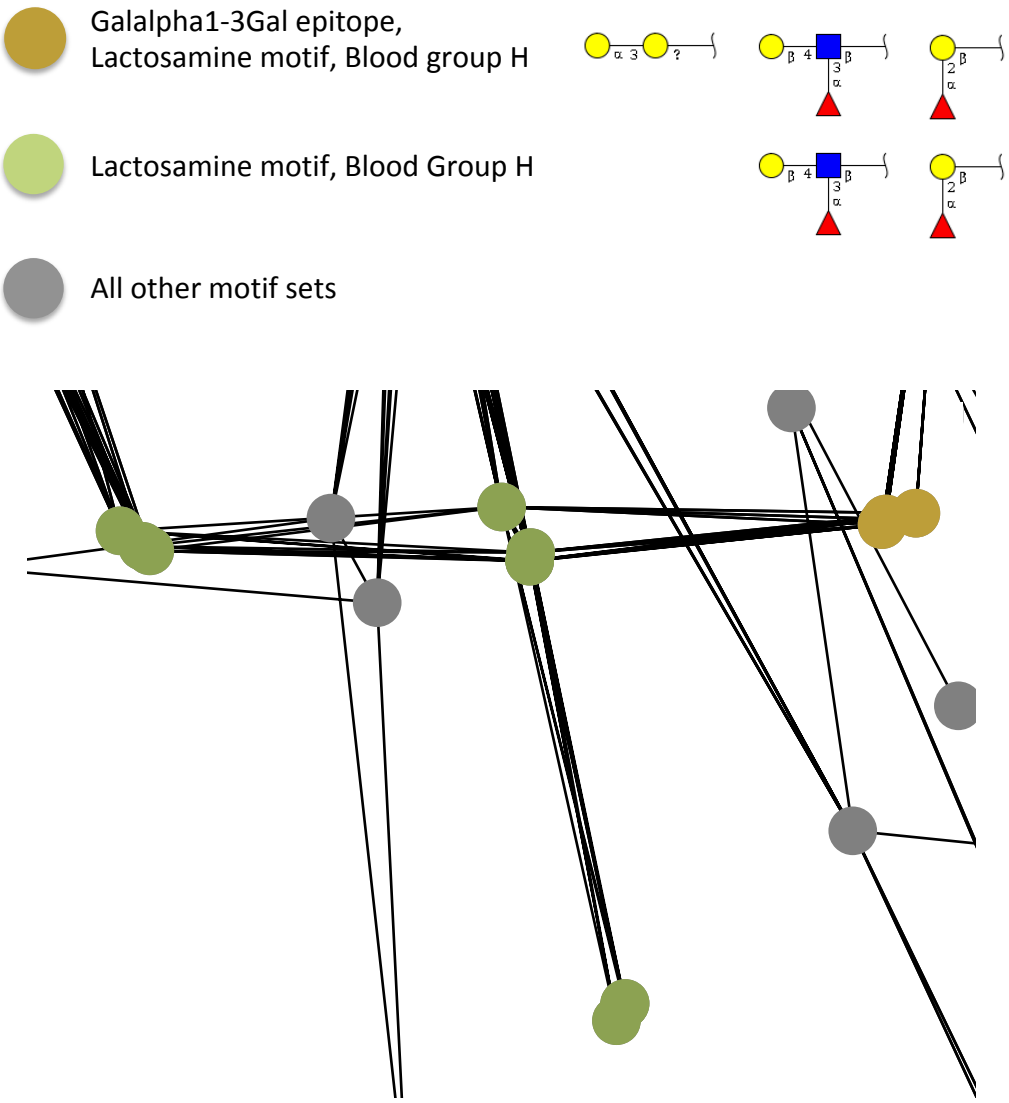


Figure 7: Glycans containing the Galalpha1-3Gal epitope, Lactosamine, and Blood group H motifs (orange nodes) are connected to glycans containing the Lactosamine, and Blood group H motifs (green nodes).

Conversely, in each of Figures 8 and 9, we highlight two groups of glycans which do not share similar motifs. In these instances, there are no direct connections between the two glycan groups, implying significant structural differences, as we would expect given the differences in motifs. Network graphs created using the

reactions list method, as opposed to the reactions set method, exhibit the same types of relationships and we omit them here.

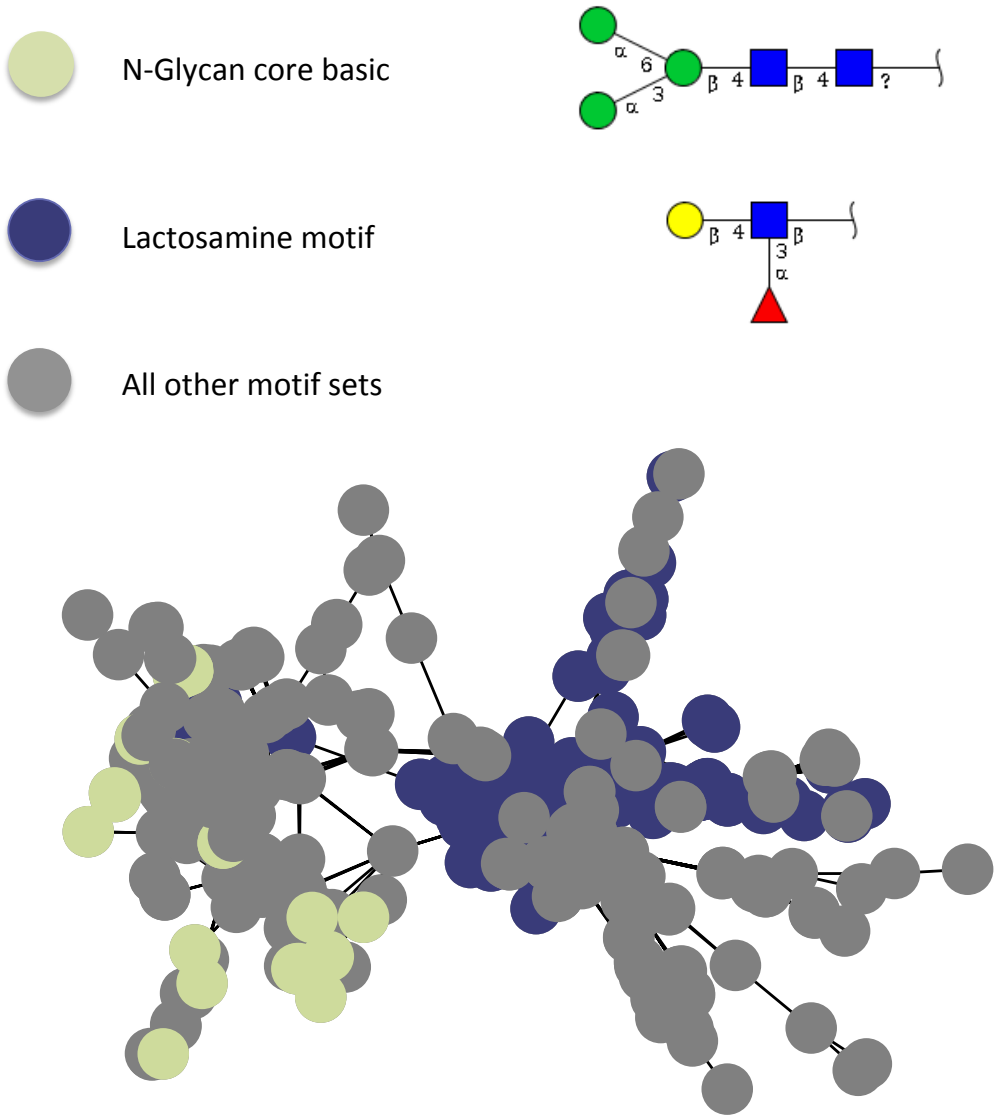


Figure 8: Glycans containing the N-Glycan core basic motif (green nodes) are not directly connected to glycans containing the Lactosamine motif (blue nodes).

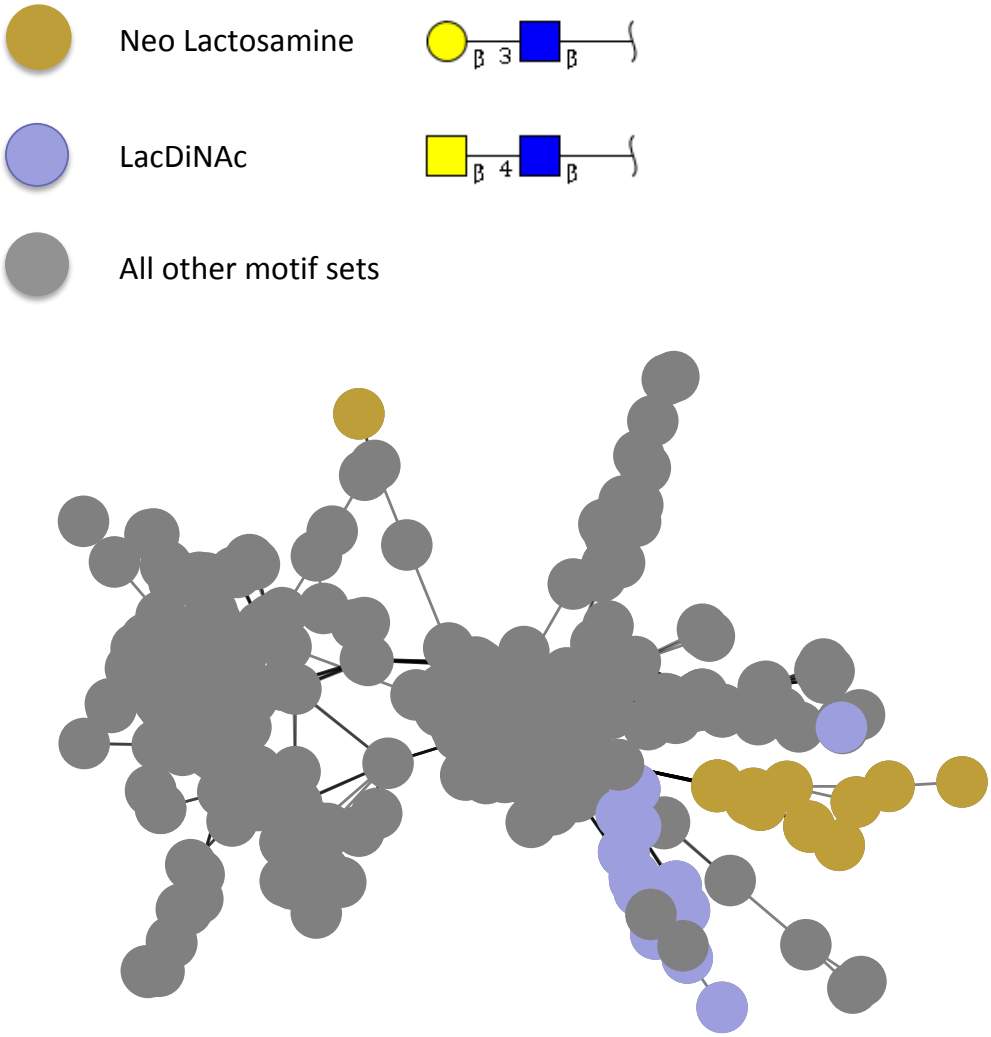


Figure 9: Glycans containing the Neo Lactosamine motif (orange nodes) are not directly connected to glycans containing the LacDiNAc motif (violet nodes).

3.2 Assortativity Coefficient

The assortativity coefficients calculated for the two alternative similarity measures are:

Reaction set method: **0.93**

Reaction list method **0.97**

The higher value for the reaction list method suggests that there is a benefit to taking into account reaction quantities, rather than just presence, when measuring glycan similarity.

3.3 UPGMA Hierarchical Clustering

Figure 10 shows a magnified section of the full tree (in cladogram form) produced using UPGMA hierarchical clustering, applied to the distance matrix obtained using the reaction list method. The 4111 terminal nodes, which represent individual glycans, are labelled with both the names of the motifs present in each glycan, and, in brackets, the glycan ID (as taken from the GlyTouCan database). The visualisation software dynamically adjusts the number of terminal nodes for which it displays labels in order to avoid clutter. We can see via the terminal labels and the branch colours that there are distinct clades of glycans which share the same motif sets, providing further evidence of the suitability of our method for determining glycan similarity. Furthermore, adjacent clades of different colours often have motifs in common, analogous to the connected nodes we observed in the glycan network.

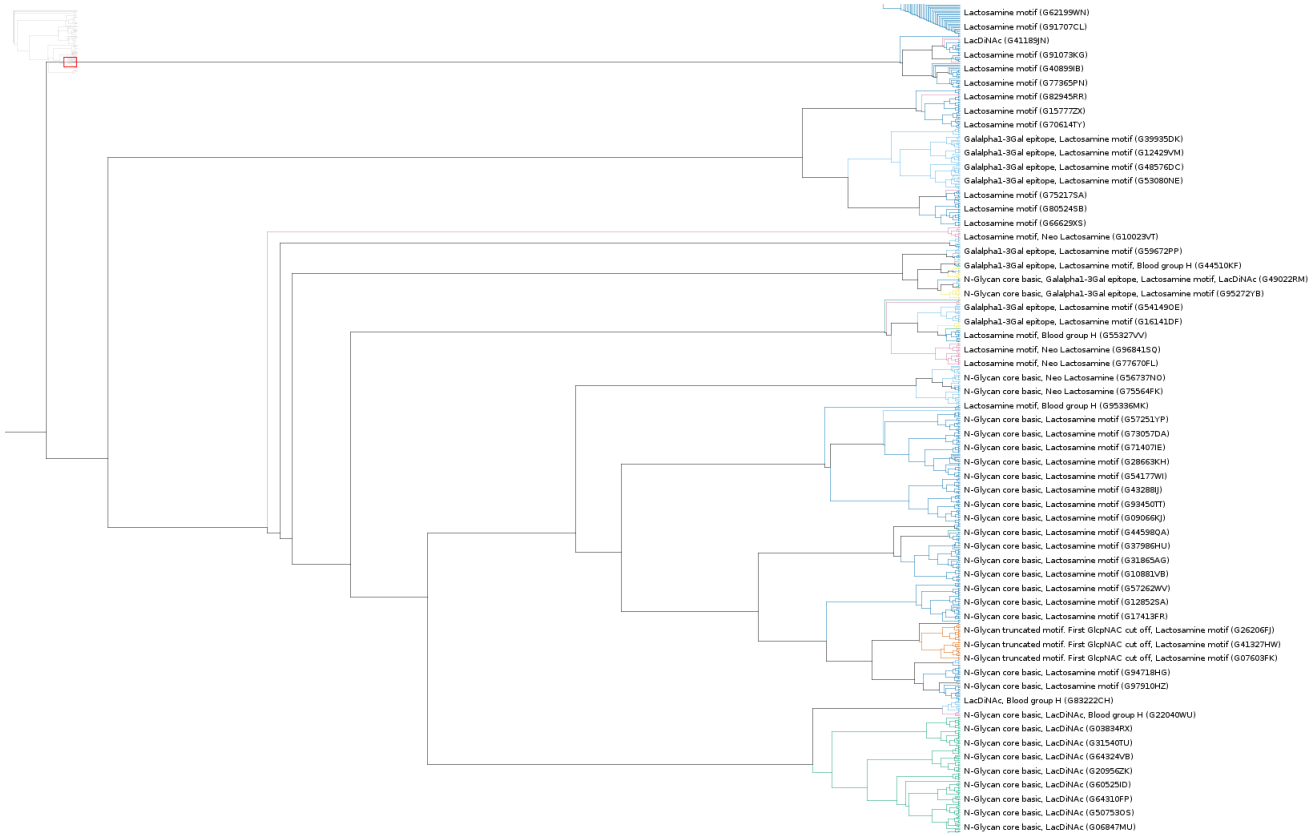


Figure 10: Full depth UPGMA tree (zoomed in to subsection). Clades of equal colour indicate clusters of glycans sharing the same motif sets.

3.3.1 Motif Discovery

A detailed inspection of the UPGMA tree reveals some features which, initially, seem to be at odds with the notion of shared motifs between similar glycans observed so far. For example, in Figure 11, a glycan with the sole motif LacDiNAc (highlighted in red) appears amongst a cluster of glycans containing only the 'N-Glycan truncated motif first GlcpNAC cut off' motif. However, looking up the structure of this glycan in the GlyTouCan database[4] reveals that this glycan does in fact contain the 'N-Glycan truncated motif first GlcpNAC cut off' motif as well as LacDiNAc, it just has not been annotated as containing it. Figure 12a shows the full glycan structure, Figure 12b shows the LacDiNAc motif, with which the glycan has been correctly annotated, and Figure 12c shows the 'N-Glycan truncated motif first GlcpNAC cut off' motif which the glycan has not been annotated as containing, but which is clearly seen in the full structure (Figure 12a). We cannot rule out though the possibility that there is some positional dependency associated with the 'N-Glycan truncated motif first GlcpNAC cut off' motif which precludes it from being declared as a motif for this particular glycan.

[4] `https://glytoucan.org/Structures/Glycans/G35645QP`

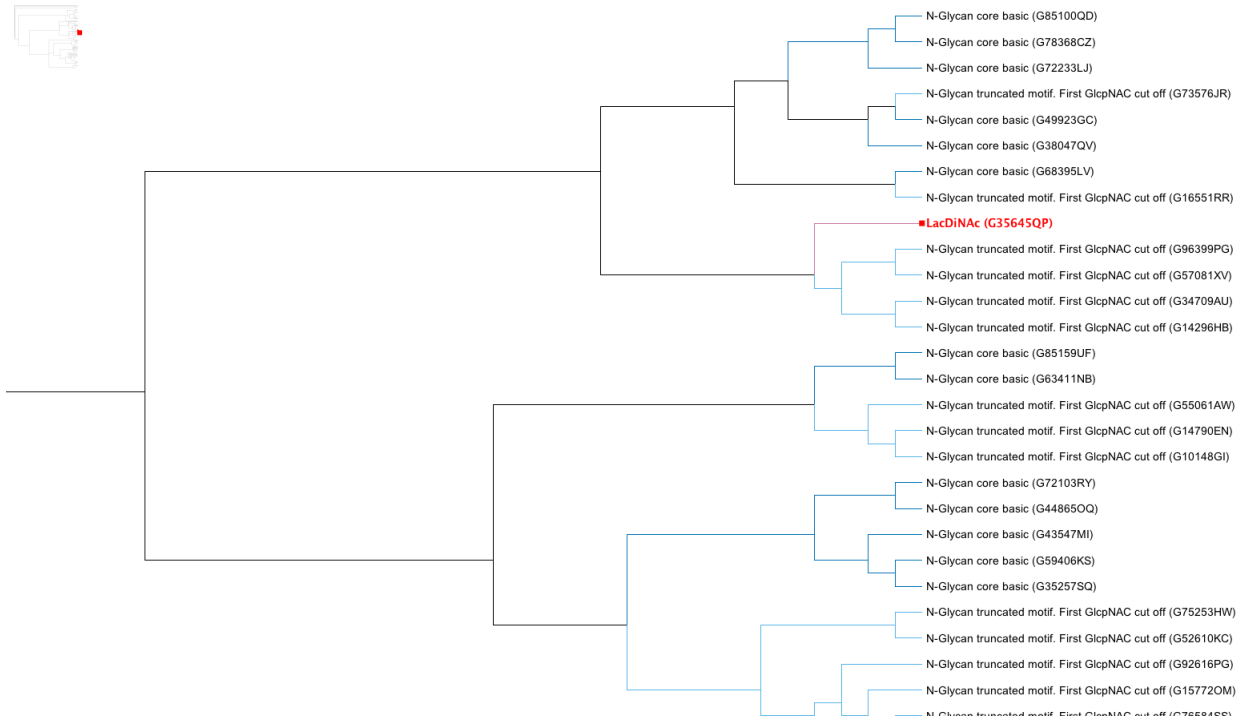
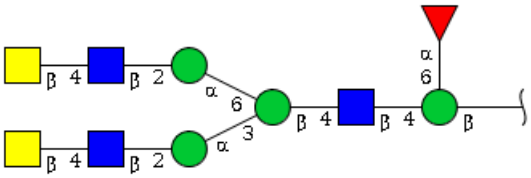
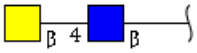


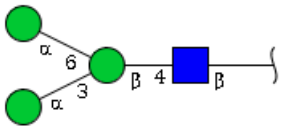
Figure 11: A glycan (highlighted in red) which has a missing motif annotation in the GlyTouCan database has been correctly clustered with glycans sharing the same motif.



(a) Actual structure of the glycan highlighted in red in Figure 11.



(b) Correctly annotated LacDiNAc motif.



(c) Missing motif annotation 'N-Glycan truncated motif First GlcpNAC cut off'

Figure 12: The actual structure of the glycan highlighted in red in Figure 11 and its present and missing motif annotations.

4 Discussion

The same-motif glycans displayed in both the connected communities of our network graphs and the clades of the UPGMA tree, as well as the discovery of the missing motif annotation for a particular glycan, are highly encouraging outcomes. These were found using just 4111 glycans from an original set of 105050 in the GlyTouCan database. Increasing the number of glycans in our analysis would open up the possibility of yet more insights of a similar nature. One straightforward way to achieve this would be to simply employ more computing resources – this study was carried out on a single moderately powerful desktop machine. Beyond that though, we currently reject a large number of glycans because of experimental uncertainty in their structure. It should be possible to modify our parsing technique to make use of some of the certain glycan reactions contained within gycans whose overall structure contains ambiguities, therefore increasing our overall coverage.

The results we present here are produced using a technique based upon characterisation of glycans in terms of the reactions necessary to construct them. Glycans exhibit a tree-like structure, which we do not take into account. Measuring distance between tree structures is inherently more complex than the technique we present here, and it would be instructive to make a comparison between this approach and one which accounts for the specific tree-like nature of glycan structure, such as glycan classification with tree kernels (Yamanishi et al. 2007).

In Section 3.2 we report and compare the network assortativity coefficients for the networks produced by our two alternative glycan similarity measurement techniques. To increase our confidence in these values we should also determine their statistical significance. The script `bootstrap_assortativity.py` is included

in the source code for this project, which uses the 'bootstrap' (Efron 1992) method to determine the means and standard deviations of the coefficients. Computation time for this script is high though, and the results were not available at the time of publication.

5 Further Work

In Sections 2.2 and 2.3 we describe querying the GlyTouCan database using SPARQL queries, and later parsing the WURCS format string in order to obtain a list of reactions for each glycan. In fact, the GlyTouCan RDF data already provides a breakdown of more fine-grained elements of the full WURCS string, which offers us the opportunity to increase the amount of glycan reactions in our dataset via the use of alternative SPARQL queries. Suggestions for exploratory SPARQL queries are given in Appendix C.

The visualisations created from our glycan networks are a valuable tool for exploring glycan similarity. The obvious communities of glycans sharing the same motif sets gives us confidence that our technique for measuring glycan similarity from their reactions is valid, and the spatial positioning offers some qualitative, if not quantitative, assessment of glycan similarity both within and between those communities of same motif set glycans. We present some static images here, however the full value of the visualisations is only realised when viewed live and interactively on a computer. We use the Python Matplotlib (Hunter 2007) package to display our networks, which allows the user to zoom in and move the image up/down/left/right. This also seems somewhat limited though, and a custom visualisation tool would enhance the value of the networks. For example, there is useful motif, structure, and similarity coefficient information which could be displayed when clicking or hovering over nodes and edges. Groups of glycans could be dynamically selected, either by mouse clicking or via a search mechanism, and compared against one another. Rendering the images in three rather than two dimensions would also be useful. Possible visualisation tools include Cytoscape

(Shannon et al. 2003) and Large Graph Layout (Adai et al. 2004).

A key factor in producing useful network visualisations is the choice of threshold value applied to the similarity matrix, in order to then reduce it down to a binary adjacency matrix. A value too high will result in only glycans with near identical structure forming connections, whilst too low will result in connections between unrelated glycans. This is not seen to be a major problem, as we are free to manipulate the networks as we please in order to maximise their utility. However it does leave open the question of where do we draw the line between similar and dissimilar glycans – further work to define this boundary would be useful.

When performing hierarchical clustering (Section 2.7, we used the UPGMA algorithm. Other algorithms could be employed here, and could potentially offer gains in terms of computation time. One example, the Bit-Pattern Biclustering (BiBit) Algorithm (Rodriguez-Baena et al. 2011), claims good performance when applied to large datasets, such as gene expression data.

The dataset we use here solely provides details of glycan structure, including motifs. A dataset of glycan samples by cell and tissue type exists (Consortium for Functional Glycomics 2010), which provides an opportunity to (for example) make comparisons between glycan structures in disease state and non-disease state cells or tissues. We provide a Python script, `glycan_scraper.py` for downloading the data from this website, however the data is in a different format (MSA) to that required by our software (WURCS). There are tools available (Chris Barnett 2016) which can convert from MSA to another format, 'linearcode', and then to WURCS, however in our tests the conversion from linearcode to WURCS was not able to deal with data in which uncertainties in the glycan structure are expressed, and consequently further development of the conversion tools is required in order

to proceed with this dataset.

In Table 1 of (Porter et al. 2009), some rules governing glycan structure in relation to motifs are given. For example, the definition given for the N-Glycan hybrid motif is "A glycan chain with a Mana1-3(Mana1-6)Manb1-4GlcNAcb1-4GlcNAcb base with GlcNAcb1-2 linked to only one of the two terminal mannose glycans (Mana1,3 or Mana1,6); the branch with the GlcNAcb1,2 can continue to grow with any glycan, but only mannose may be present on the other mannose". An extension of the tools we present here could be used to validate these rules, for example by modelling glycans which break the rules and then searching for real examples of high similarity.

6 Conclusions

Measuring similarity between glycans is challenging due to their complex tree-like structure and the high number of both constituent monosaccharides and the types of glycosidic links between them. We have demonstrated encouraging results based upon a novel technique of determining glycan similarity which avoids the need to carry out complex comparisons of tree structures, allowing us to explore similarities both within and between communities of glycans sharing the same or similar sets of motifs. Our technique enables the identification of previously unlabelled motifs in glycans, and we provide several suggestions for advancing this research further. Of particular interest is the possibility of using the cell and tissue-specific glycan dataset in order to broaden the scope of the research into identifying glycan structures associated with disease states.

We present network graph and cladogram images offering a qualitative validation of the technique, and would encourage the adaptation of the computer code provided such that our results might be visualised and explored further using more advanced imaging software.

Some relatively straightforward modifications to the method of querying the glycan database, and the use of more powerful computing resources, open up the possibility of vastly increasing the number of glycans analysed, and we provide the necessary computer script to determine the statistical significance of our quantitative results in order to further validate the technique.

References

Adai, A. T., Date, S. V., Wieland, S. & Marcotte, E. M. (2004), 'Lgl: creating a map of protein function with an algorithm for visualizing very large biological networks', *Journal of molecular biology* **340**(1), 179–190.

Albelda, S. M., Smith, C. W. & Ward, P. (1994), 'Adhesion molecules and inflammatory injury.', *The FASEB Journal* **8**(8), 504–512.

Aoki, K. F., Mamitsuka, H., Akutsu, T. & Kanehisa, M. (2004), 'A score matrix to reveal the hidden links in glycans', *Bioinformatics* **21**(8), 1457–1463.

Aoki, K. F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S. & Kanehisa, M. (2004), 'Kcam (kegg carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains', *Nucleic acids research* **32**(suppl_2), W267–W272.

Bennun, S. V., Yarema, K. J., Betenbaugh, M. J. & Krambeck, F. J. (2013), 'Integration of the transcriptome and glycome for identification of glycan cell signatures', *PLOS Computational Biology* **9**(1), 1–18.
URL: *https://doi.org/10.1371/journal.pcbi.1002813*

Bucior, I. & Burger, M. M. (2004), 'Carbohydrate–carbohydrate interactions in cell recognition', *Current opinion in structural biology* **14**(5), 631–637.

Campbell, M. P., Ranzinger, R., Lütteke, T., Mariethoz, J., Hayes, C. A., Zhang, J., Akune, Y., Aoki-Kinoshita, K. F., Damerell, D., Carta, G. et al. (2014), 'Toolboxes for a standardised and systematic study of glycans', *BMC bioinformatics* **15**(1), S9.

Chris Barnett (2016), 'Glycome analytics platform', `https://bitbucket.org/scientificomputing/glycome-analytics-platform`. [Online; accessed 4-September-2018].

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et al. (2009), 'Biopython: freely available python tools for computational molecular biology and bioinformatics', *Bioinformatics* **25**(11), 1422–1423.

Consortium for Functional Glycomics (2010), 'Glycan profiling', `http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling-new.jsp/`. [Online; accessed 4-September-2018].

Cossart, P. & Sansonetti, P. J. (2004), 'Bacterial invasion: the paradigms of enteroinvasive pathogens', *Science* **304**(5668), 242–248.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

Fruchterman, T. M. & Reingold, E. M. (1991), 'Graph drawing by force-directed placement', *Software: Practice and experience* **21**(11), 1129–1164.

Fuster, M. M. & Esko, J. D. (2005), 'The sweet and sour of cancer: glycans as novel therapeutic targets', *Nature Reviews Cancer* **5**(7), 526.

Gronau, I. & Moran, S. (2007), 'Optimal implementations of upgma and other common clustering algorithms', *Information Processing Letters* **104**(6), 205–210.

Hagberg, A., Swart, P. & S Chult, D. (2008), Exploring network structure, dy-

namics, and function using networkx, Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Han, M. V. & Zmasek, C. M. (2009), 'phyloxml: Xml for evolutionary biology and comparative genomics', *BMC bioinformatics* **10**(1), 356.

Hosoda, M., Akune, Y. & Aoki-Kinoshita, K. F. (2017), 'Development and application of an algorithm to compute weighted multiple glycan alignments', *Bioinformatics* **33**(9), 1317–1323.

Hunter, J. D. (2007), 'Matplotlib: A 2d graphics environment', *Computing In Science & Engineering* **9**(3), 90–95.

Kawano, S., Hashimoto, K., Miyama, T., Goto, S. & Kanehisa, M. (2005), 'Prediction of glycan structures from gene expression data based on glycosyltransferase reactions', *Bioinformatics* **21**(21), 3976–3982.
URL: *http://dx.doi.org/10.1093/bioinformatics/bti666*

Matsubara, M., Aoki-Kinoshita, K. F., Aoki, N. P., Yamada, I. & Narimatsu, H. (2017), 'Wurcs 2.0 update to encapsulate ambiguous carbohydrate structures', *Journal of chemical information and modeling* **57**(4), 632–637.

McNaught, A. D. & McNaught, A. D. (1997), *Compendium of chemical terminology*, Vol. 1669, Blackwell Science Oxford.

Newman, M. E. (2003), 'Mixing patterns in networks', *Physical Review E* **67**(2), 026126.

Porter, A., Yue, T., Heeringa, L., Day, S., Suh, E. & Haab, B. B. (2009), 'A

motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins', *Glycobiology* **20**(3), 369–380.

Rodriguez-Baena, D. S., Perez-Pulido, A. J. & Aguilar-Ruiz, J. S. (2011), 'A biclustering algorithm for extracting bit-patterns from binary datasets', *Bioinformatics* **27**(19), 2738–2745.

Rosa, C. & Reinhold, V. N. (2002), Functional post-translational proteomics approach to study the role of n-glycans in the development of caenorhabditis elegans, *in* 'Biochem. Soc. Symp', Vol. 69, pp. 1–21.

Sacks, D. & Kamhawi, S. (2001), 'Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis', *Annual reviews in microbiology* **55**(1), 453–483.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003), 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome research* **13**(11), 2498–2504.

Song, E.-H., Shang, J. & Ratner, D. (2012), 9.08 - polysaccharides, *in* K. Matyjaszewski & M. Möller, eds, 'Polymer Science: A Comprehensive Reference', Elsevier, Amsterdam, pp. 137 – 155.
URL: *http://www.sciencedirect.com/science/article/pii/B9780444533494002466*

Staden, R. (1979), 'A strategy of dna sequencing employing computer programs', *Nucleic acids research* **6**(7), 2601–2610.

Tong, L., Baskaran, G., Jones, M. B., Rhee, J. K. & Yarema, K. J. (2003), 'Glycosylation changes as markers for the diagnosis and treatment of human disease', *Biotechnology and Genetic Engineering Reviews* **20**(1), 199–246.

Ueda, N., Aoki-Kinoshita, K. F., Yamaguchi, A., Akutsu, T. & Mamitsuka, H. (2005), 'A probabilistic model for mining labeled ordered trees: capturing patterns in carbohydrate sugar chains', *IEEE Transactions on Knowledge and Data Engineering* **17**(8), 1051–1064.

Von Der Lieth, C.-W., Bohne-Lang, A., Lohmann, K. K. & Frank, M. (2004), 'Bioinformatics for glycomics: status, methods, requirements and perspectives', *Briefings in Bioinformatics* **5**(2), 164–178.

W3C SPARQL Working Group (2013), 'Sparql 1.1 overview', `https://www.w3.org/TR/sparql11-overview/`. [Online; accessed 4-September-2018].

W3C Working Group (2014), 'Rdf 1.1 primer', `https://www.w3.org/TR/rdf11-primer/`. [Online; accessed 4-September-2018].

Yamanishi, Y., Bach, F. & Vert, J.-P. (2007), 'Glycan classification with tree kernels', *Bioinformatics* **23**(10), 1211–1216.
URL: *http://dx.doi.org/10.1093/bioinformatics/btm090*

Appendices

A Software Dependencies

Table 3 lists the packages necessary to run the python code used to produce the results described in this paper. The program code itself can be cloned into the current directory by typing the following command in a command window (assuming the Git[5] application is already installed):

```
git clone https://github.com/ImperialCollegeLondon/glycans.git
```

Package Name	Version
biopython	1.72
fontconfig	2.12.6
glypy	0.12.2
html5lib	1.0.1
httplib2	0.11.3
matplotlib	2.2.2
networkx	2.1
numpy	1.14.3
numpy-base	1.14.5
pandas	0.23.2
python	3.6.6
rdflib	4.2.2
scipy	1.1.0
seaborn	0.8.1
sparqlwrapper	1.8.0
urllib3	1.23

Table 3: Python software packages necessary to run the program code associated with this project.

[5] `https://git-scm.com/`

B Glycan Databases and Structure Formats

Table 4 lists some glycan databases, the number of glycans they contain, and the glycan structure formats they provide.

Database Name	No. glycans	Formats
KEGG Glycan[a]	504	KCF
GlyConnect[b]	3768	IUPAC
UniCarbKB[c]	3238	note[d]
GlyTouCan[e]	105050	WURCS, GlycoCT, IUPAC Condensed, IUPAC Extended

Table 4: Glycan databases, and the structure formats they provide.

[a] `https://www.genome.jp/dbget-bin/www_bfind_sub?mode=bfind&max_hit=1000&locale=en&serv=gn&dbkey=glycan&keywords=*&page=1`
[b] `https://glyconnect.expasy.org/browser/structures/2259`
[c] `http://www.unicarbkb.org/`
[d] Website not working at time of publication
[e] `https://glytoucan.org/`

Different glycan structure formats exist. The number of glycans in the GlyTouCan database for each format it supports are shown in Table 5.

Format Name	Number of Glycans	Comments
GlycoCT	45438	Verbose, no longer supported
IUPAC	14517	
IUPAC condensed	73329	
IUPAC extended	73329	
WURCS	105050	Most recently published

Table 5: Glycan counts in the GlyTouCan database for different glycan structure formats.

Additional glycan resources can be found at the following URLs:

- `https://biosciencedbc.jp/en/db-link/d09-dblink`
- `http://www.functionalglycomics.org/static/consortium/links.shtml`
- `https://biosciencedbc.jp/en/db-link/d09-dblink`

C SPARQL queries

SPARQL queries may be run programmatically, or against the GlyTouCan database at the following 'SPARQL Endpoint' (a web query interface):

`https://ts.glytoucan.org/sparql`

The following SPARQL query returns all Glycan IDs, WURCS format strings, and motif IDs. Results for glycans with multiple motifs will be spread across multiple lines:

```
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX glytoucan:  <http://www.glytoucan.org/glyco/owl/glytoucan#>

SELECT DISTINCT ?Saccharide ?PrimaryId ?Sequence
                ?Motif ?MotifPrimaryId
FROM <http://rdf.glytoucan.org/core>
FROM <http://rdf.glytoucan.org/sequence/wurcs>
FROM <http://rdf.glytoucan.org/motif>
WHERE {
  ?Saccharide glytoucan:has_primary_id ?PrimaryId .
  ?Saccharide glycan:has_glycosequence ?GlycoSequence .
  ?GlycoSequence glycan:has_sequence ?Sequence .
  ?GlycoSequence glycan:in_carbohydrate_format
    glycan:carbohydrate_format_wurcs.
  OPTIONAL { ?Saccharide glycan:has_motif ?Motif .
             ?Motif glytoucan:has_primary_id ?MotifPrimaryId } .
}
ORDER BY ?PrimaryId
```

This SPARQL query returns all Motif IDs and motif labels:

```
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX glytoucan:  <http://www.glytoucan.org/glyco/owl/glytoucan#>

SELECT DISTINCT ?Motif ?MotifPrimaryId ?MotifLabel
FROM <http://rdf.glytoucan.org/core>
FROM <http://rdf.glytoucan.org/sequence/wurcs>
FROM <http://rdf.glytoucan.org/motif>
WHERE {
  ?Saccharide glycan:has_motif ?Motif .
  ?Motif glytoucan:has_primary_id ?MotifPrimaryId.
  ?Motif rdfs:label ?MotifLabel
}
ORDER BY ?MotifPrimaryId
```

This SPARQL query is useful for exploring the glycan properties available in the GlyTouCan database, and returns a list of the distinct RDF properties within the full set of triples:

```
SELECT DISTINCT ?p
FROM <http://rdf.glytoucan.org/core>
FROM <http://rdf.glytoucan.org/sequence/wurcs>
FROM <http://rdf.glytoucan.org/motif>
WHERE {
    ?s ?p ?o
}
```

A selection of results from the query above which could be useful for constructing new exploratory SPARQL queries are listed overleaf.

http://purl.jp/bio/12/glyco/glycan#has_glycosequence
http://purl.jp/bio/12/glyco/glycan#has_resource_entry
http://purl.jp/bio/12/glyco/glycan#has_sequence
http://purl.jp/bio/12/glyco/glycan#in_carbohydrate_format
http://purl.jp/bio/12/glyco/glycan#in_glycan_database
http://purl.jp/bio/12/glyco/glycan#has_motif
http://www.glycoinfo.org/glyco/owl/wurcs#LIN_count
http://www.glycoinfo.org/glyco/owl/wurcs#RES_count
http://www.glycoinfo.org/glyco/owl/wurcs#has_GLIP
http://www.glycoinfo.org/glyco/owl/wurcs#has_GLIPS
http://www.glycoinfo.org/glyco/owl/wurcs#has_LIN
http://www.glycoinfo.org/glyco/owl/wurcs#has_MAP_position
http://www.glycoinfo.org/glyco/owl/wurcs#has_RES
http://www.glycoinfo.org/glyco/owl/wurcs#has_SC_position
http://www.glycoinfo.org/glyco/owl/wurcs#has_basetype
http://www.glycoinfo.org/glyco/owl/wurcs#has_direction
http://www.glycoinfo.org/glyco/owl/wurcs#has_monosaccharide
http://www.glycoinfo.org/glyco/owl/wurcs#has_rep_max
http://www.glycoinfo.org/glyco/owl/wurcs#has_rep_min
http://www.glycoinfo.org/glyco/owl/wurcs#has_root_RES
http://www.glycoinfo.org/glyco/owl/wurcs#has_uniqueRES
http://www.glycoinfo.org/glyco/owl/wurcs#is_fuzzy
http://www.glycoinfo.org/glyco/owl/wurcs#is_monosaccharide
http://www.glycoinfo.org/glyco/owl/wurcs#is_repeat
http://www.glycoinfo.org/glyco/owl/wurcs#is_uniqueRES
http://www.glycoinfo.org/glyco/owl/wurcs#uniqueRES_count
http://www.glycoinfo.org/glyco/owl/wurcs#has_modification_prob_lower
http://www.glycoinfo.org/glyco/owl/wurcs#has_modification_prob_upper
http://www.glycoinfo.org/glyco/owl/wurcs#has_MAP
http://www.glycoinfo.org/glyco/owl/wurcs#has_acceptor
http://www.glycoinfo.org/glyco/owl/wurcs#has_donor
http://www.glytoucan.org/glyco/owl/glytoucan#has_primary_id
http://www.glytoucan.org/glyco/owl/glytoucan#date_registered
http://www.glytoucan.org/glyco/owl/glytoucan#contributor
http://www.glycoinfo.org/glyco/owl/wurcs#has_root_GRES
http://www.glycoinfo.org/glyco/owl/wurcs#has_GRES
http://www.glycoinfo.org/glyco/owl/wurcs#is_acceptor_of
http://www.glycoinfo.org/glyco/owl/wurcs#has_acceptor_position
http://www.glycoinfo.org/glyco/owl/wurcs#has_donor_position
http://www.glycoinfo.org/glyco/owl/wurcs#is_donor_of
http://www.glycoinfo.org/glyco/owl/wurcs#count_RES
http://www.glycoinfo.org/glyco/owl/wurcs#count_uniqueRES
http://www.glycoinfo.org/glyco/owl/wurcs#count_LIN
http://www.glytoucan.org/glyco/owl/glytoucan#is_reducing_end

D Illustrative WURCS Format String Example

Figure 13 (overleaf) illustrates the components of a WURCS format string, describing glycan structure. The corresponding brreakdown in terms of the biological reactions necessary for glycan synthesis is also shown.

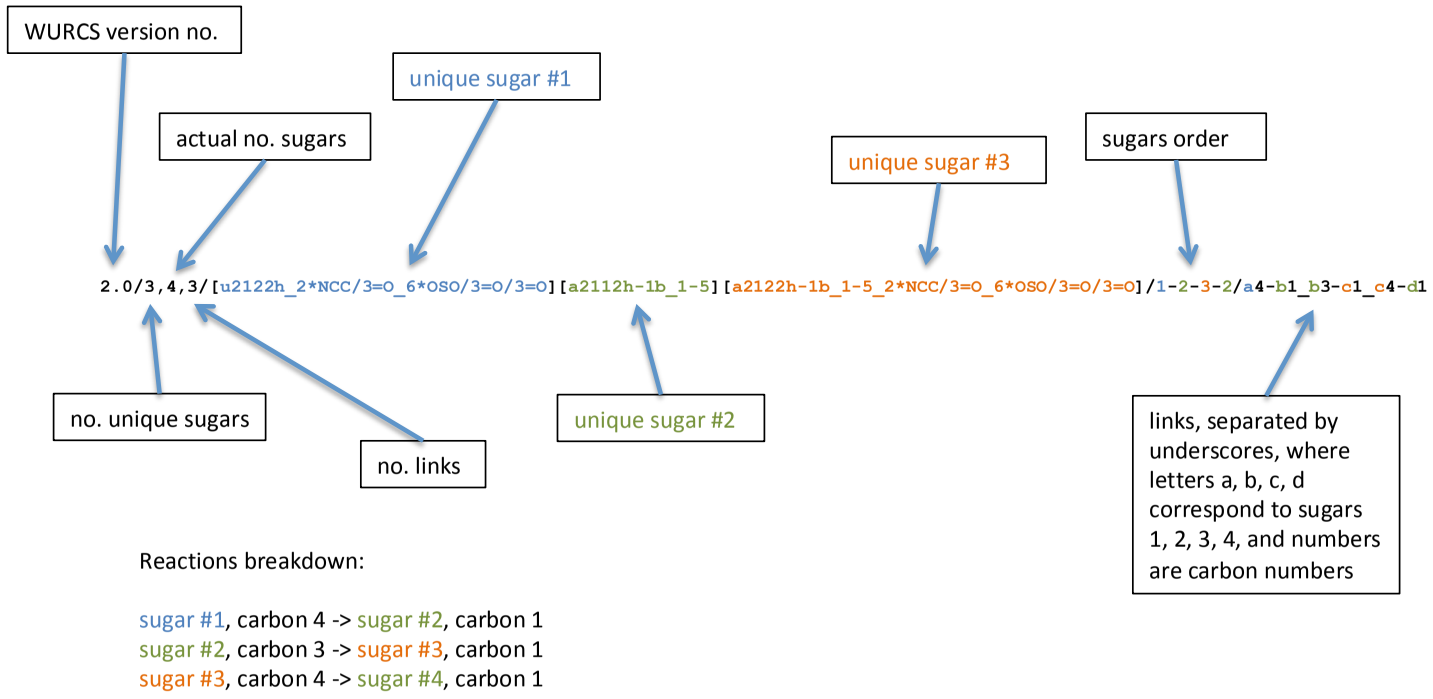


Figure 13: An example of glycan structure expressed in the WURCS format, with a corresponding breakdown in terms of child sugar, glycosidic link, and parent sugar. This breakdown is what we use to characterise glycans in terms of the biological reactions necessary for their synthesis.