Exploring the Landscape of Glycan Structure

MSc Bioinformatics and Theoretical Systems Biology Project

Steven Hargreaves

CID 01453122

Imperial College London
Department of Life Sciences

Abstract

contributions

- identify mis-laballed motifs

- we can determine glycan structure similarity

– without needing to model the precise tree structure

– so could also predict structure of unknown glycans from GE data or enzyme abundance

remove URLs from refs

Acknowledgements

Many thanks to John Pinney for his supervision and guidance during this project, and Suhail Islam for technical assistance.

Contents

Abbreviations

Glossary

1 Introduction

1.1 Glycans

Monosaccharides are carbohydrates which cannot be further hydrolyzed 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). In both cases, 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.

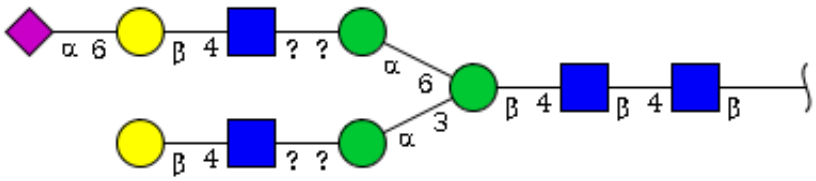


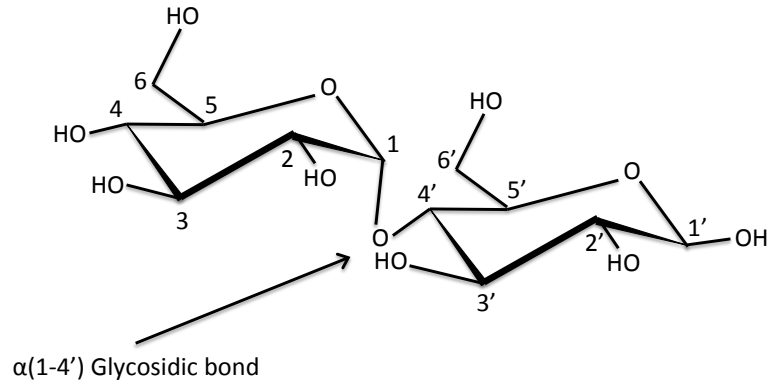
Figure 1: An example glycan diagram. The differently coloured and shaped nodes represent different monosaccharides, and the labelled edges represent the glycosidic links between them (image taken from the GlyTouCan repository[1]). Glycosidic links are characterised by the anomericity (α or β) of carbon 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.

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

ALSO SHOW MONOSACCHARIDE KEY?

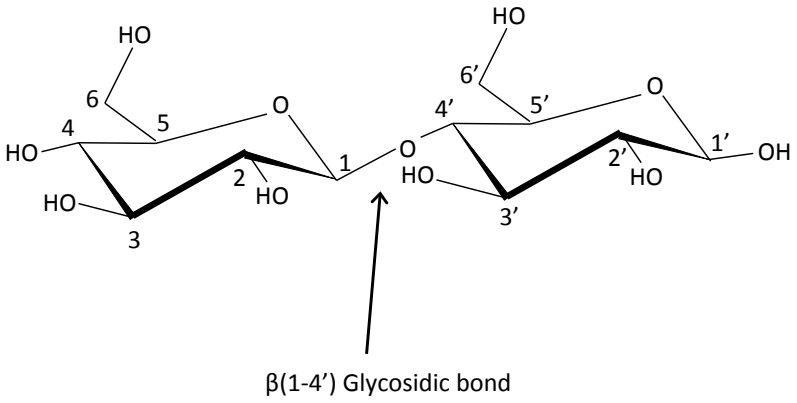
When dissolved in an aqueous environment, the monosaccharides (or sugars) of which glycans are comprised, primarily exist in the cyclic (or hemiacetal) form. In this form, the sugars are further characterised by their anomeric configurations – either the α-anomer, with an axial OH-group at carbon 1, 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 carbon on the second sugar. This can be seen more clearly in Figure 1, where, for example, the XXX (purple diamond) is glycosidically linked via its α-anomer carbon 1 to carbon 6 of the XXX (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).

When glycans are linked with proteins or lipids, they are characterised as being either N-linked, in which a monosaccharide is attached to a nitrogen atom of a protein (Taylor & Drickamer 2011), or O-linked, in which a glycosidic oxygen links the glycoside to the aglycone or reducing end sugar (LOOK FOR AN ALTERNATIVE, CITABLE EXPLANATION OF THIS, E.G. IN ALBERTS OR THE BIG BIOCHEM BOOK).



[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).

1.2 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). There are a total of 105050 distinct glycan structures present in the GlyTouCan[2] database. 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 (with diagrams) is available via the footnote link[3].

STATE / SHOW THE THREE MAJOR MOTIF TYPES (see wiki / glytoucan)

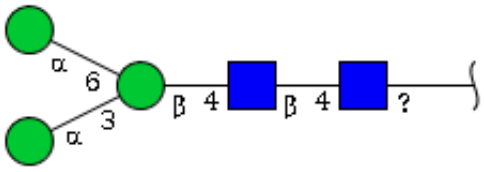


Figure 3: The 'N-Glycan core basic' glycan motif, which can also be seen as part of the structure of the glycan shown in Figure 1.

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

1.3 Glycan Function

Evidence has been found relating glycans to numerous important biological functions, including immune reaction, protein stabilization, and cell signalling (Bucior & Burger 2004), tumour progression (Fuster & Esko 2005), and embryogenesis (Rosa & Reinhold 2002). 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).

1.4 Glycan Analysis

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. 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 to-

gether 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).

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 (REF), 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), and glycan comparison (Aoki, Mamitsuka, Akutsu & Kanehisa 2004, Aoki, Yamaguchi, Ueda, Akutsu, Mamitsuka, Goto & Kanehisa 2004) MAYBE A BIT MORE DETAIL ABOUT THESE TECHNIQUES AND RESULTS.

— describe/list existing databases?

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 catalyze the reactions necessary for synthesising specific glycans. By building a library of the bond formation patterns of GT reactions, and determining a co-occurence 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 distance 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 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 / cladogram ???, from which we compare the groups of similar glycans found using the two alternative methods - THIS NEEDS RE-THINKING - SAY MORE ABOUT REACTIONS, INCLUDE A DIAGRAM LIKE Figure 4.

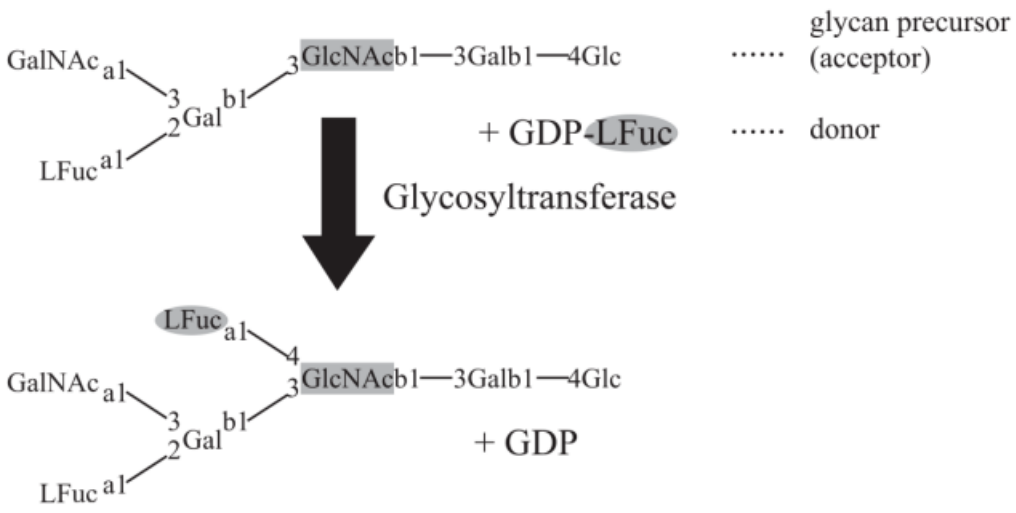


Figure 4: NEED TO RE-DRAW THIS

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.

1.6 MAKE SURE YOU'VE GOT ALL OF JOHN'S REFS IN. AND ALSO THE MOTIF PAPER cwp187.pdf

2 Method

Figure 5 illustrates, at a high level, the methods used in this research. In brief, we:

- Extract glycan structure and motif information from a public Glycan database (GlyTouCan)
- Save the data locally
- Parse the structure data, extracting the collection of reactions necessary to construct each glycan
- Use the Jaccard distance measure to determine similarity between all pairs of glycans in terms of their reactions collections
- Convert the similarity measurements to binary similar / dissimilar by application of a threshold
- Create, visualise, and explore a network of similar glycans
- Construct, visualise and explore a cladogram via the use of hierarchical clustering

In the following sub-sections we describe each of these steps in more detail.

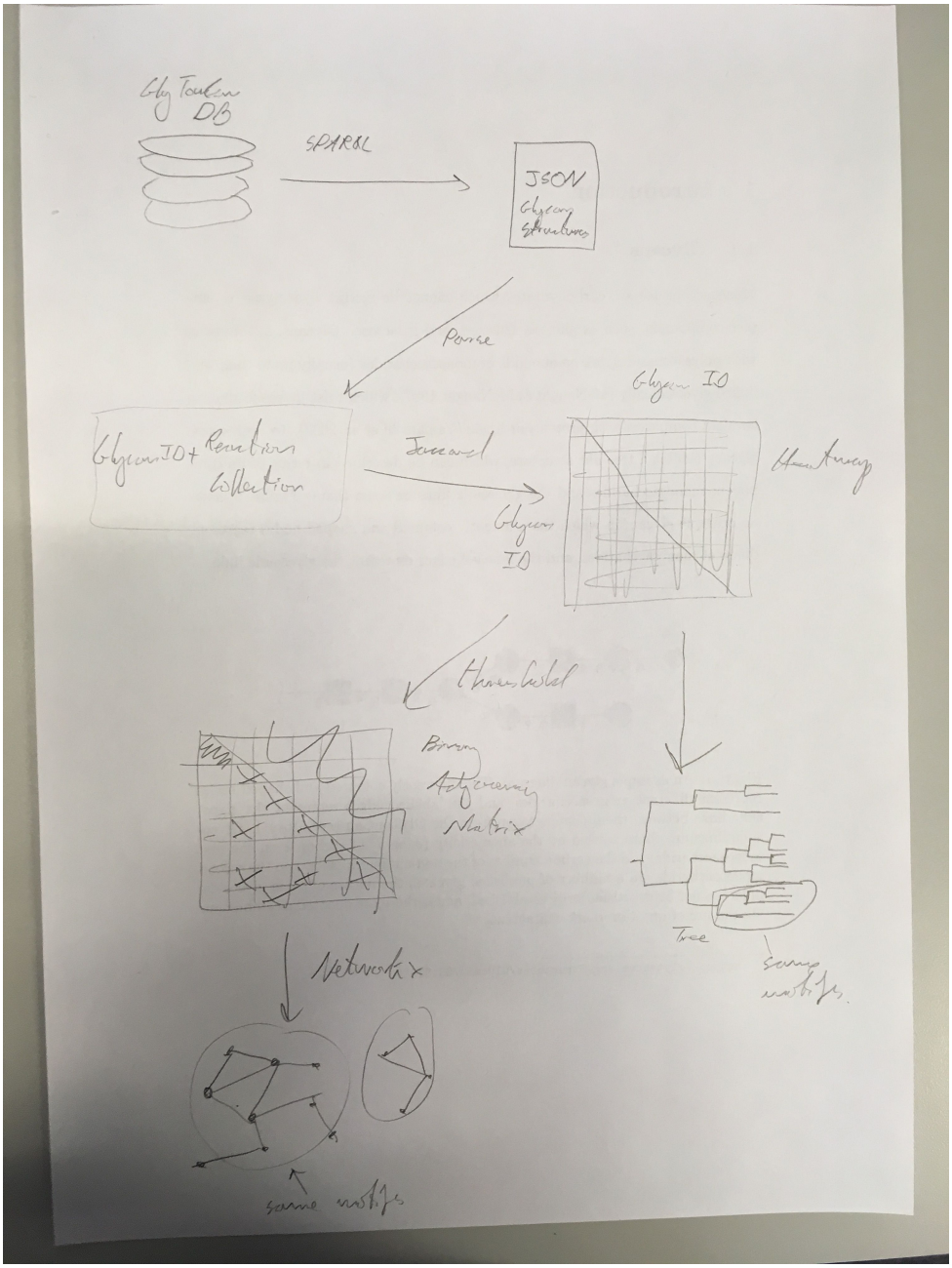


Figure 5: Method flowchart

2.1 Glycan Databases

These exist:

KEGG Glycan database[4] contains 504 glycans. Formats: KCF ?

GlyConnect[5] contains 3768 glycans. Formats: IUPAC.

UniCarbKB[6] somewhat confusing interface. 3238 GlycoSuite Structures, 899 Glycoproteins.

GlyTouCan[7] 105050 glycans. Formats: WURCS, GlycoCT, IUPAC Condensed, IUPAC Extended. Also motifs.

Note - GlycoCT format seems to be 'dead' - was an EBI thing but funding ran out. Subsumed by `http://unicarb-db.expasy.org/` ?

More glycan resources here `https://biosciencedbc.jp/en/db-link/d09-dblink`, here `http://www.functionalglycomics.org/static/consortium/links.shtml`, and here `https://biosciencedbc.jp/en/db-link/d09-dblink`.

We chose GlyTouCan, because most glycans, most formats, motif info, easy to query.

Different glycan structure formats exist. The number of glycans in GlyTouCan

[4] `https://www.genome.jp/dbget-bin/www_bfind_sub?mode=bfind&max_hit=1000&locale=en&serv=gn&dbkey=glycan&keywords=*&page=1`

[5] `https://glyconnect.expasy.org/browser/structures/2259`

[6] `http://www.unicarbkb.org/`

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

for each format are shown in Table 1.

- GlycoCT

– Verbose

– Past EBI project, no longer supported / funded

– Possibly subsumed by unicarb-db.expasy.org

- IUPAC

- IUPAC condensed

- IUPAC extended

- WURCS

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

We chose GlyTouCan as DB.

- 105050 glycans, 61 motifs, RDF database, SPARQL endpoint.

– give some RDF & SPARL background

– explain how your queries work, e.g. the need to query across multiple graphs to get motif data.

— Glytoucan SPARQL endpoint uses GlycoRDF (I think) - ontology info here `https://github.com/ReneRanzinger/GlycoRDF`

Queried databse for list of glycan ID / Sequence (WURCS string) / Motif ID - all saved to a json file

- there can be multiple Motif IDs per glycan ID, so describe how we handle that

2.2 Extracting Reactions from Glycan Structure Data

The WURCS spec

- v1.0

– extremely terse, not easily human readable

- v2.0

– can handle ???

Parse the WURCS string for each glycan to get collection of reactions

- Problems: see python code (uncertainties etc.)

- Ignore those we can't parse (state this as a limitation in conclusions, and something for further work)

2.3 Calculating Distances Between Glycan Reaction Collections

Use Jaccard (both methods) to create similarity matrix of glycans in terms of reaction collection similarity

- see handwritten notes (photo in email)

– Jaccard set method Figure 6

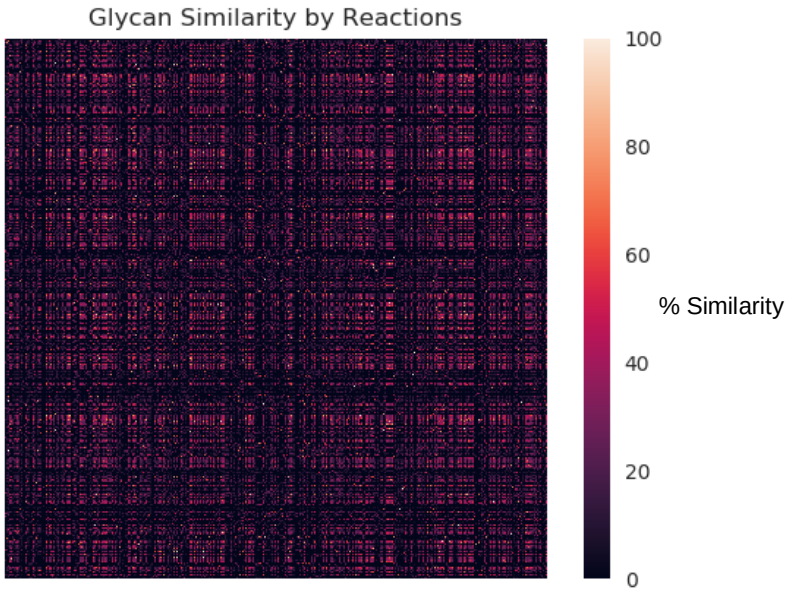


Figure 6: Glycan Reaction Sets Heatmap CROP IMAGE. Each axis is a list of Glycans IDs, not shown here because there are too many to display.

– Jaccard list method

- make sure None - None matches (i.e. no reactions?) are marked as 0 rather than 100% similarity

- describe algorithm, e.g. the use of sets etc.

- matrix is saved, not re-calculated every time

(check python code - do we ignore ¿1 motifs? at what point are zero connection nodes taken out? in the most connected components?)

2.4 Binary Adjacency Matrix

Apply threshold to make a binary distinction between similar / dissimilar

2.5 Glycan Network Creation

Treat this as an adjacency matrix, and, using a column-labelled dataframe, create a (networkx) graph

Choose suitable layout

assortativity coeff

get X most connected components

Redo using reaction quantities (modified Jaccard)

- ensure new network has same no. of edges

– use weighted edges, order, trim

- * generate some new network images? (low priority, might just show more compact clusters)

2.6 Hierarchical Clustering

UPGMA hierarchical clustering

- how does it work - apply motif sets to terminal node names

- colour the nodes & branches

– describe how

— e.g. manipulating phyloxml to produce coloured trees in archeopteryx

- does it highlight possible mistakes? e.g. the isolated Galalpha1 somewhere in tree for method one (set)

- images: circular if poss, otherwise cladogram cos branch lengths meaningless

2.7 Reproducibility

The computer program code used to produce the results described in this paper 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.

3 Results

(Discuss them as you present them)

- Walk through the network

– Figs 5 to 10: Glycan (un)connected components

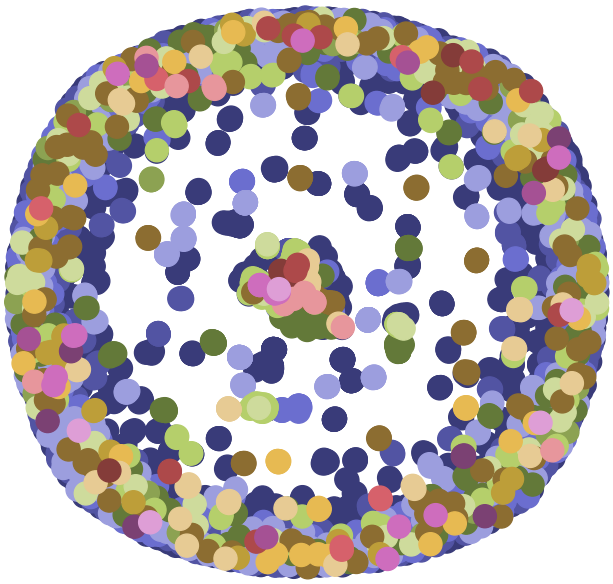


Figure 7: Full network of glycans, with similarity determined using the reaction set method. CROP.

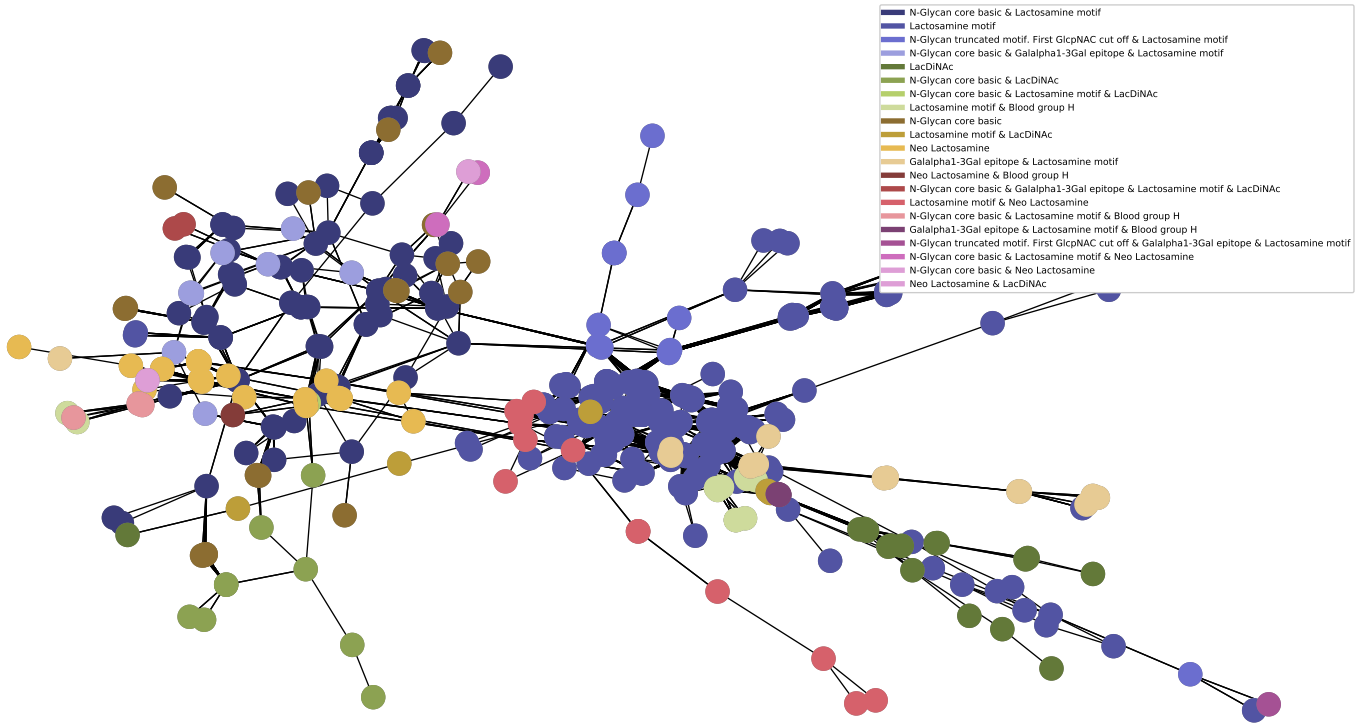


Figure 8: Most connected component of a network of glycans, with similarity determined using the reaction set method. CROP and TURN THIS SIDEWAYS.

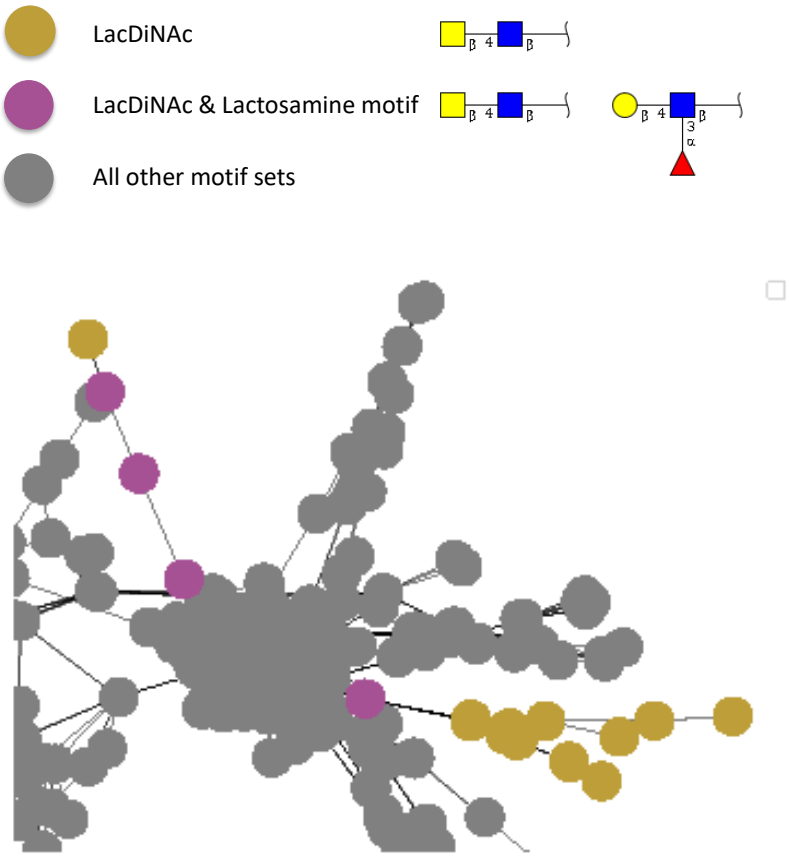


Figure 9: LacDiNAc vs Lactosamine motif,LacDiNAc. CROP.

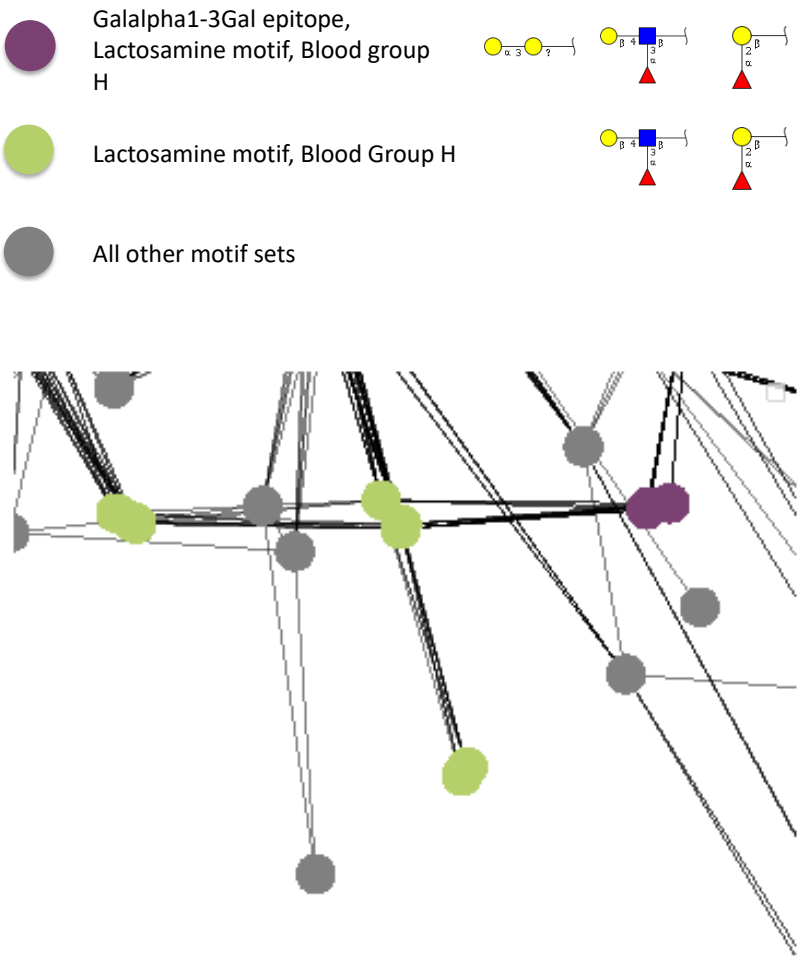


Figure 10: Galalpha1-3Gal epitope, Lactosamine motif, Blood group H vs Lactosamine motif, Blood group H. CROP.

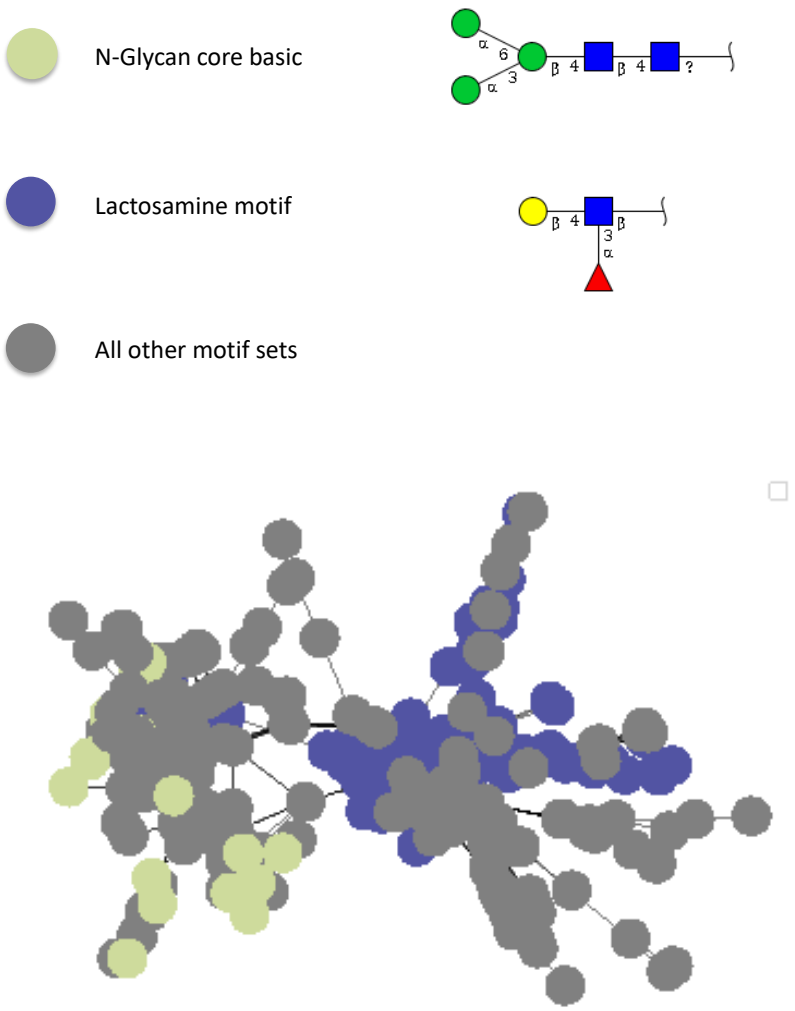


Figure 11: N-Glycan core basic vs Lactosamine motif. CROP.

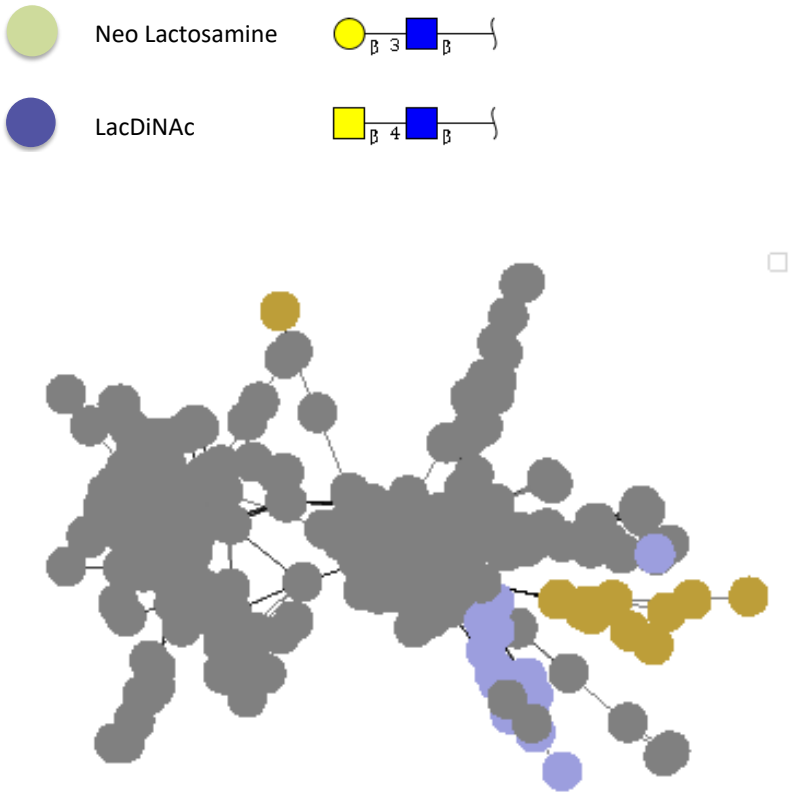


Figure 12: Neo Lactosamine vs LacDiNAc. CROP.

— Several obvious communities apparent

— In moving from community A to B, we gain/lose motif X — The fact that

nodes with the same motif sets are spaced out reflects the fact that the glycans have more structure than the motifs alone, and our method is showing that, so that's added value.

- Assortativity coeff for set vs list method

– statistically significant? — either do this calc, or write about it in Discussion

—- see if there's an easy way to get a number from networkx. Otherwise, bootstrap / jacknife etc.

- explore the UPGMA tree

– use different depths / zooms

— Fig 11: full tree

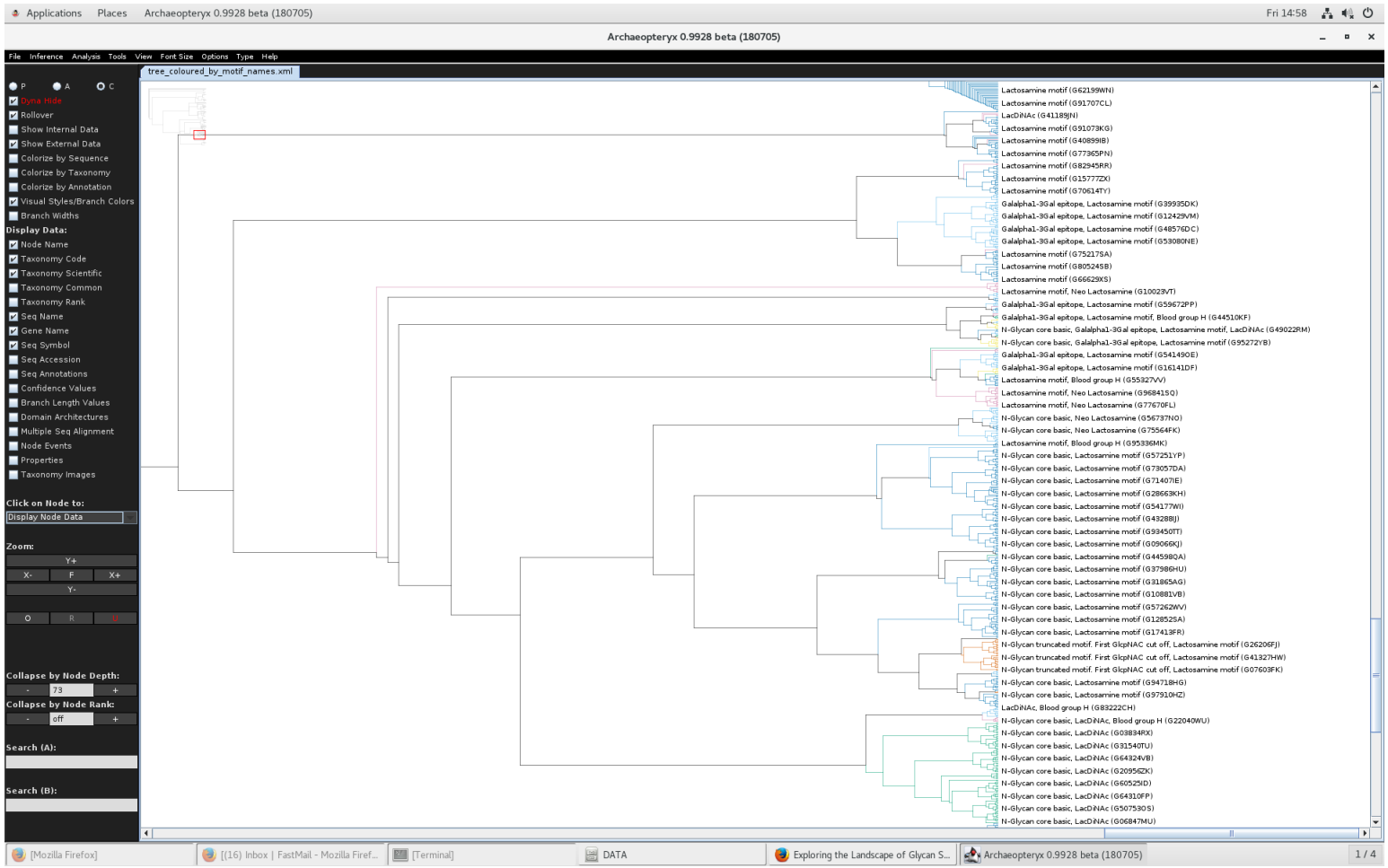


Figure 13: Full UPGMA tree (zoomed in). CROP and TURN THIS SIDEWAYS.

— Fig 12: zoomed-in tree

– describe the clades, relate them to the network if poss

— therefore make a comparison between jaccard + network and UPGMA tree methods

—- Are there similarities between tree and network? (network is only a vis tool, the tree is the real deal)

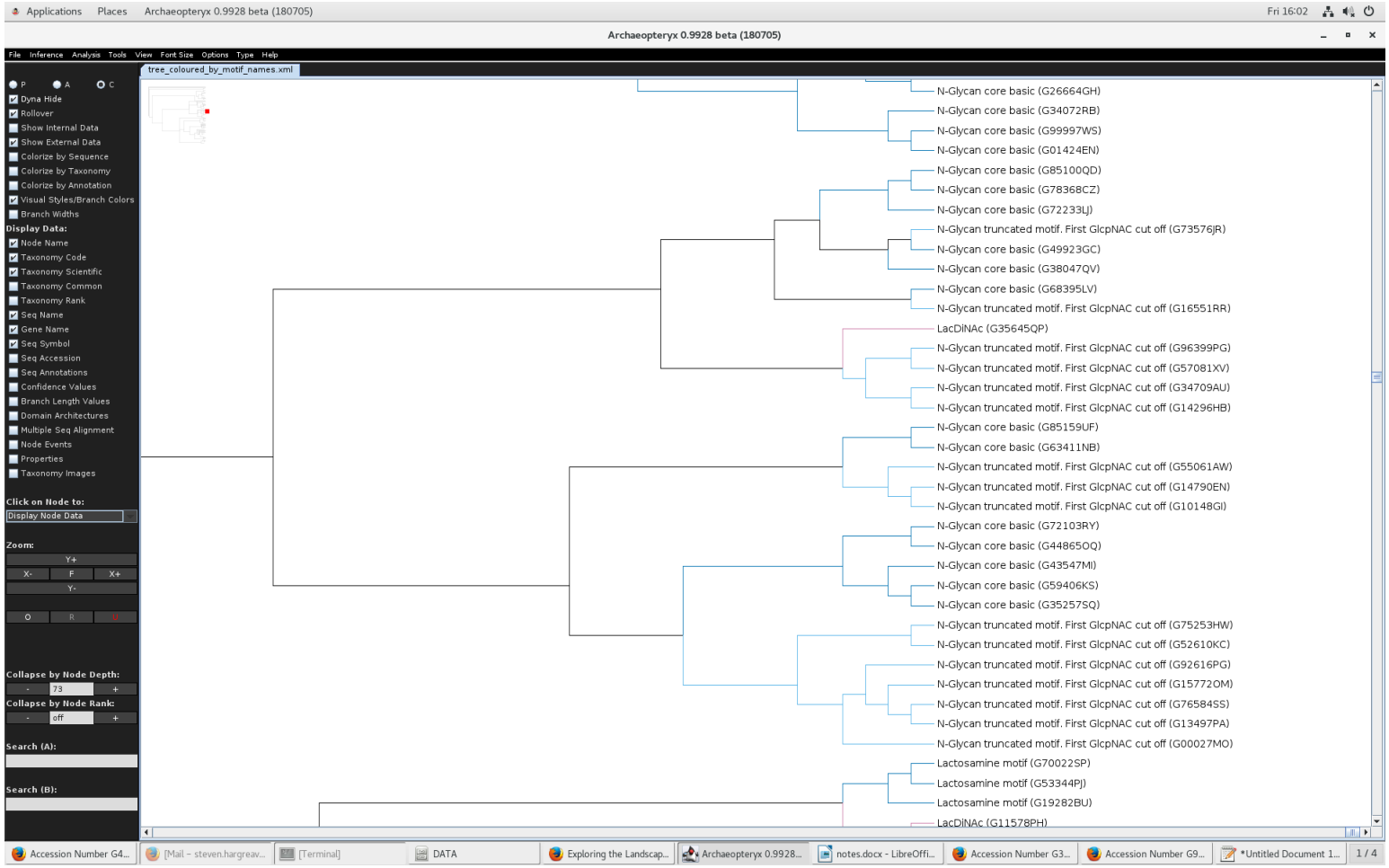
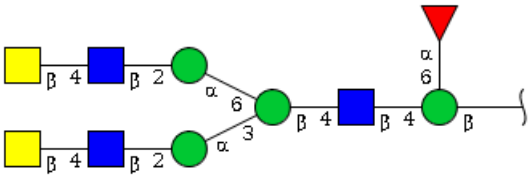
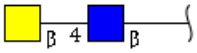


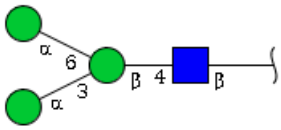
Figure 14: Rogue LacDiNAc. CROP and TURN THIS SIDEWAYS.



(a) Rogue LacDiNAc structure. CROP



(b) LacDiNAc motif



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

Figure 15: Rogue LacDiNAc and motifs.

4 Discussion

(On the bigger picture)

Should have calculated statistical significance of assortativity coeffs. Could use bootstrap etc.

5 Further Work

Can use more advanced SPARQL query to get reaction breakdown (so should increase both number of glycans and accuracy)

Bi-clustering

Kernel methods

- New dataset, samples by cell / tissue
- details on what we would do – Work already done (web scraping for data)
– Format conversion problems
— conversion software breaks for linearcode to WURCS (can't handle uncertainties)

Compare full structures, not just reactions

Compare sets of structures

Use neighbour joining or K-means instead of UPGMA (can you say why that might be useful?)

Other glycan databases?

KEGG?

Use LGL, organic cytoscape for vis

chi-squared to compare two (quantified) reaction profiles

6 Conclusions

Appendices

A Software Dependencies

Table 2 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[8] 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 2: Python software packages necessary to run the program code associated with this project.

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

References

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.

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

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

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*

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

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.

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*

Taylor, M. E. & Drickamer, K. (2011), *Introduction to glycobiology*, Oxford university press.

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.

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*