Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Masters in Biomedical Engineering, Hiren Joshi
born in London, United Kingdom

Oral-examination: TBA

Automatic validation of glycan sequences in distributed databases

Referees:
Prof. Dr. Sabine. Strahl
Prof. Dr. Roland. Eils

Automatic validation of glycan sequences in distributed databases

Hiren Joshi

First supervisor: Prof. Sabine Strahl

The study of glycosylation is an emerging field that aims to understand the structure, synthesis and function of the commonly found molecules known as glycans. This integrative area of study covers many sub-domains such as chemistry, biology and informatics.

Structurally, glycans are aggregate molecules, composed of a set of monosaccharides linked together through glycosidic linkages to form oligo or polysaccharides. Glycans are commonly found conjugated to other molecules such as proteins and lipids, or as free molecules in their own right.

The biosynthesis of glycans is not template-based — a simple reading of the genome or the proteome will not yield any information as to the total set of glycans in a system (also known as the glycome). In fact, the biosynthesis of glycans follows a complex pathway that depends on many factors, resulting in unique glycosylation profiles for sub-systems and tissues.

In order to overcome the inability to predict the complete glycome, efforts have been made to develop glycan databases that collect the observed glycans for a diverse set of systems. These databases are critically important to the development of the field, as they represent the knowledge base; the ability to query and search this information is an integral part of the toolset that researchers use to make further discoveries.

The open model for database curation is attractive for the low-maintenance, decentralised nature of data collection. However, the use of open curation has implications for the quality of the data, which must be maintained through the implementation of validation procedures. This thesis is an investigation into the use of automatic validation techniques within openly curated databases.

One possible application for this technology is within the EUROCarbDB project — a Europe wide effort to build an openly curated database that allows for easy deposition of glycan structures as well as the associated primary data used to establish each structure.

The main aims of this thesis are to establish a method for automatically validating structures deposited into a glycomic database in order to maintain its quality; to ascertain whether it is possible to use these validation methodologies to obtain an estimate of the size of the human glycome; and to promote the distributed nature of the EUROCarbDB project, so as to provide a testing ground for the validation techniques.

The main achievements in this thesis are many. First and foremost, I developed an algorithm for validating structures against pathway data. Also, I investigated the ability of the pathway and enzymatic data to explain the synthesis of human structures. Anomalous enzymatic data was identified for further examination, and the size of the human glycome was successfully estimated. Finally, a networking layer was established within the EUROCarbDB, necessitating the future use of the validation algorithms.

Automatische Validierung von Glykanstrukturen in verteilten Datenbanken

(Automatic validation of glycan sequences in distributed databases)

Hiren Joshi

Betreuer: Prof. Sabine Strahl

Der Forschungsbereich Glykobiologie integriert wissenschaftliche Erkenntnisse aus der Chemie, Biologie und der Informatik mit dem Ziel, die Struktur, Synthese und die Funktionen von Glykanen zu untersuchen.

Glykane bestehen aus Monosacchariden, welche mittels glykosidischen Verbindungen zu Oligo- oder Polysacchariden verkettet werden. Man findet sie meistens gebunden an Proteine oder Lipide oder als freie Moleküle. Die Struktur und Synthese der Gesamtheit aller Glykane (das Glykom) ist nicht direkt im Genom oder Proteom kodiert, sondern erfolgt durch eine Enzymkaskade und wird durch Faktoren wie Expressionsprofil der Enzyme und Verfügbarkeit von Monosaccharid-Substraten beeinflusst.

Das Resultat ist ein heterogenes Glykosylierungsprofil in den verschiedenen Gewebetypen und anatomischen Strukturen.

Datenbanken dienen der Archivierung von funktionellen und strukturellen Informationen über Glykane und ermöglichen deren Analyse mittels rechnerischen Ansätzen. Diese Datenbanken, besonders die Qualität der darin enthaltenen Daten, sind von zentraler Bedeutung für die Forschungsgemeinschaft.

Für die Sammlung von dezentral produzierten Forschungsergebnissen in Datenbanken ist ein offenes Kurationsmodell attraktiv, da die Erzeuger der Daten diese selbständig in die Datenbasis einfügen. Um eine hohe Qualität und Konformität der Daten zu gewährleisten ist die Entwicklung automatisierter Methoden erforderlich.

Die Zielsetzung dieser Arbeit ist die Entwicklung von Methoden zur Qualitätskontrolle in Datenbanken mittels automatischen Ansätzen zur Validierung von Glykanstrukturen. Eine Anwendung der entwickelten Verfahren zur Herleitung der Größe des menschlichen Glykoms und Integration der entwickelten Technologien findet sich im Rahmen des EUROCarbDB Projektes. Dieses Projekt befasst sich mit der Etablierung einer offen kuratierten Datenbank zur Archivierung von Glykanstrukturen und assoziierten strukturaufklärenden Messdaten.

Die Ergebnisse umfassen die Entwicklung eines Algorithmus zur Validierung von Glykanstrukturen mittels Informationen über deren Biosynthese. Nicht verifizierbare Enzymdaten und Glykanstrukturen wurden identifiziert und eine Größenbestimmung des humanen Glykoms durchgeführt. Des Weiteren wurde EUROCarbDB mit Netzwerk-Funktionalität ausgestattet, welche die in dieser Arbeit entwickelten Methoden anwendet.

Contents

LIST OF FIGURES

List of Tables

Abbreviations

DKFZ	German Cancer Research Center
EBI	European Bioinformatics Institute
API	Application Programming Interface
HTTP	HyperText Transfer Protocol
IT	Information Technology
MVC	Model View Controller
ORM	Object Relational Mapping
SQL	Structured Query Language
JDBC	Java Database Connectivity Application Programming Interface (API)
CRUD	Create, Read, Update and Destroy
OOP	Object Oriented Programming
UniProt	Universal Protein resource
NCBI	National Center for Biotechnology Information
MeSH	Medical Subject Headings
MEDLINE	Medical Literature Analysis and Retrieval System Online
REST	REpresentational State Transfer
XML-RPC	XML Remote Procedure Call
WSDL	Web Services Description Language
PKI	Public Key Infrastructure
DFS	Distributed File System

GPL	GNU Public License
GID	Global Identifier
LID	Local Identifier
DBMS	DataBase Management System
CCRC	Complex Carbohydrate Research Center
CCSD	Complex Carbohydrate Structural Database
CDG	Congenital Disorder of Glycosylation
ER	endoplasmic reticulum
FTP	File Transfer Protocol
Fuc	Fucose
Gal	Galactose
Glc	Glucose
GalNAc	N-Acetyl-D-galactosamine
GlcNAc	N-Acetyl-D-glucosamine
Man	Mannose
NeuAc	N-Glycolyl Neuraminic acid
NeuGc	N-Acetyl Neuraminic acid
Xyl	Xylose
GlcA	Glucuronic acid
ERAD	Endoplasmic Reticulum Associated Protein Degradation
GPI	Glycosylphosphatidylinositol
GAG	Glycosaminoglycans
GSL	Glycosphingolipid
HTML	Hypertext Markup Language
HPLC	High Performance Liquid Chromotography

IUPAC	International Union of Pure and Applied Chemistry
KEGG	Kyoto Encyclopaedia of Genes and Glycans
NMR	Nuclear Magnetic Resonance spectroscopy
MS	Mass Spectrometry
PNG	Portable Network Graphic
SVG	Scalable Vector Graphics
UML	Unified Modelling Language
XHTML	eXtensible Hypertext Markup Language (HTML)
XML	eXtensible Markup Language

Symbols

NeuAc	◆
Xyl	☆
GlcN	◪
Gal	🟡
Fuc	🔺
Glc	🔵
NeuGc	◇
IdoA	⬙
GalN	◪
GalNAc	🟨
Man	🟢
GlcNAc	🟦

Nomenclature

Addend For any given addition operation, there are two addends: $addend + addend = result$.

EUROCarbDB A European Union funded database project to create a distributed evidence-based repository for glycan data. More information can be found at the home page (http://www.eurocarbdb.org/).

Glyco-active enzyme An enzyme that is involved in the synthesis of glycans, as well as their degradation.

Glycogene A gene that codes for an enzyme involved in glycosylation — i.e. a glyco-active enzyme.

Glycome The complete set of glycan structures found in a particular biological system.

Graph Both a data structure and a network representation. A graph consists of nodes and edges, where the edges connect the nodes together. Trees are a special form of graph, where there are no cycles along any path in the graph.

JDBC Java Database Connectivity API

Leaf A node on a graph whereby the node is only directly connected (adjacent) to one other node.

Path A path can be formed along a graph by keeping track of the nodes and edges that have been visited along a traversal, or walk along the graph.

Root Arbitrarily defined node in a tree, whereby only one parent edge is connected to it. In tree structures, the root node is the parent node where all child nodes are linked. A traversal from any node in a tree via the parent node will always lead to the root.

Sibling A sibling node is a node that shares the same parent as another node. So for example, nodes A and B may be siblings if they share a common parent node C.

SOAP Protocol for eXtensible Markup Language (XML) based web services

Tranche Distributed file system for scientific data

Tree A specialised graph that is defined as having no cycles along any path in the graph.

Preface

The study of glycosylation is an emerging field that aims to understand the structure, synthesis and function of the commonly found molecules known as glycans. This integrative area of study covers many sub-domains such as chemistry, biology and informatics.

Structurally, glycans are aggregate molecules, composed of a set of monosaccharides linked together through glycosidic linkages to form oligo or polysaccharides. Glycans are commonly found conjugated to other molecules such as proteins and lipids, or as free molecules in their own right.

The biosynthesis of glycans is not template-based — a simple reading of the genome or the proteome will not yield any information as to the total set of glycans in a system (also known as the glycome). In fact, the biosynthesis of glycans follows a complex pathway that depends on many factors, resulting in unique glycosylation profiles for sub-systems and tissues.

In order to overcome the inability to predict the complete glycome, efforts have been made to develop glycan databases that collect the observed glycans for a diverse set of systems. These databases are critically important to the development of the field, as they represent the knowledge base; the ability to query and search this information is an integral part of the toolset that researchers use to make further discoveries.

The open model for database curation is attractive for the low-maintenance, decentralised nature of data collection. However, the use of open curation has implications for the quality of the data, which must be maintained through the implementation of validation procedures. This thesis is an investigation into the use of automatic validation techniques within openly curated databases.

One possible application for this technology is within the EUROCarbDB project — a Europe wide effort to build an openly curated database that allows for easy deposition of glycan structures as well as the associated primary data used to establish each structure.

The main aims of this thesis are to establish a method for automatically validating structures deposited into a glycomic database in order to maintain its quality; to ascertain whether it is possible to use these validation methodologies to obtain an estimate of the size of the human glycome; and to promote the distributed nature of the EUROCarbDB project, so as to provide a testing ground for the validation techniques.

The main achievements in this thesis are many. First and foremost, I developed an algorithm

for validating structures against pathway data. Also, I investigated the ability of the pathway and enzymatic data to explain the synthesis of human structures. Anomalous enzymatic data was identified for further examination, and the size of the human glycome was successfully estimated. Finally, a networking layer was established within the EUROCarbDB, necessitating the future use of the validation algorithms.

Chapter 1

Glycobiology and glycobioinformatics

Glycosylation is the most commonly observed — and most inherently diverse — post translational modification in biological systems, and is also the most inherently diverse. The regulation, function and structure of the glycans that result from glycosylation are the subject of the field of glycobiology. Glycans are synthesised as a product of genes and proteins. It is through proteins — specifically the glycan-related enzymes — that glycan synthesis is primarily effected. Regulation of synthesis occurs in a non-template, systems-based manner [1], involving many factors ranging from gene expression to substrate availability. Coupled with the branching nature of glycans, this results in the diversity of glycan products. The study of glycosylation is a difficult area due to the complexity of the structures being examined.

1.1 Glycobiology

1.1.1 History

The study of glycobiology has historically been approached from two sides-the study of carbohydrate chemistry, and the study of biochemistry. An early pioneer in the field of carbohydrate chemistry was Emil Fisher, who studied simple carbohydrates as early as 1891. The field of glyocobiology grew during the 20th century (reviewed in Roseman [2]) with improvements in analytical techniques. The term glycobiology was coined in 1988 by Rademacher, Parekh and Dwek [3].

1.1.2 Structure

The term glycan can often be used interchangeably with carbohydrate. Monosaccharides, disaccharides, oligosaccharides and polysaccharide are all carbohydrates. Both oligosaccharide and polysaccharide have well defined meanings with respect to the structure of sugars. Polysaccharides are large polymers, while oligosaccharides are smaller glycans. Monosaccharides are generally found as cyclic structures, with 3 to 10 carbon atoms within the ring. See Figure 1.1 for examples of some common monosaccharides.

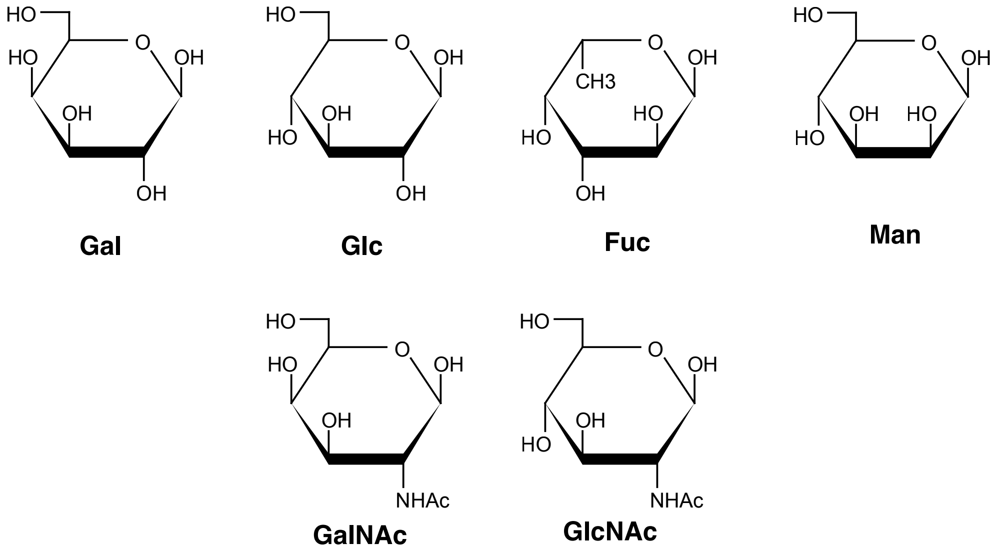


Figure 1.1: Hayworth projections of some common human monosaccharides.

Glycans are found either conjugated to other molecules, or as free molecules in their own right. Glycans can be found attached to proteins and lipids, or can form parts of structures such as the Glycosylphosphatidylinositol (GPI) anchors and Glycosaminoglycans (GAG)s. Glycans are attached to proteins through an asparagine, serine or threonine amino acid. Lipids are joined to the glycan through a ceramide unit. GPI anchors are attached via phosphatidylinositol, whereas GAGs are N-linked (keratin sulfate) or O-linked (other GAGs) to proteins to form molecules known as proteoglycans. The attachment point on the non-carbohydrate component of the molecule can be used as a classifier for the glycans: N-linked, O-linked, GAG, Glycosphingolipid (GSL) or GPI anchor. This classifier also gives hints as to the structural motifs found on the attached structures, since each point has different families of structures attached.

1.1.3 Function

Elucidation of the relationship between glycan structure and function is one of the main goals of glycomic analysis. By clarifying the ways that structure affects cellular function, the medical applications of glycans can be revealed. Currently, relatively little is known about the exact function of various structural features in glycans, and the mechanisms by which function is effected are varied. Generally, the main mode of action is through protein-carbohydrate interaction — specifically through the interaction with lectins, a family of carbohydrate-active proteins known to bind glycan epitopes.

Recognition Glycan recognition is the most general mechanism of inferring a function for a glycan. Lectins contain carbohydrate-recognition domains which serve as selectors of particular glycan epitopes or structural features. By either having non-binding functional domains, or forming complexes with other functional molecules, different behaviour can be associated with recognition of the glycan sequence. Receptor epitopes can range from being very specific — on the level of ring substitution positions (such as haemagglutinin binding specifically to NeuAc linked to a residue on a 6-substituted residue [4]) — to non-specific (such as binding to a single monosaccharide regardless of the neighbouring residues). In a protein synthesis quality control role, Endoplasmic Reticulum Associated Protein Degradation (ERAD) is controlled by mannose trimming [5], requiring the protein conformation-dependent loss of mannose units to ensure that the protein is degraded at the right point. Glycan tagging is a specific feature of the generalised recognition mechanism which uses the properties of the conjugated molecule to conditionally modify the appearance of glycan receptor epitopes.

Tagging Tagging of glycoconjugates occurs when the properties of the conjugated molecule direct the addition of glycan tags onto the molecule. In the case of glycoproteins, the protein conformation can be used to determine whether a particular glycan epitope is available for lectin binding. For example, an improperly folded glycoprotein in the calnexin-calreticulin cycle is recognised by a glycosyltransferase, causing it to bind to calnexin or calreticulin [5]. The conformation of the conjugated molecule triggers the addition of a monosaccharide residue, which causes the glycoconjugate to be sorted by a glycan tag-specific lectin. Similarly, a GlcNAc-phosphate transferase is used to tag lysosomal hydrolyses for transport away from the Golgi. Following tagging, a GlcNAc glycosidase is applied, leaving a mannose-6-phosphate epitope which is recognised by a lectin that sorts the hydrolyses to the endosomes.

Protection Certain glycoproteins such as LAMP-1 and LAMP-2 are highly glycosylated, resulting in the formation of an almost contiguous carbohydrate coat that protects the proteins against degradation by lysosomal proteases [5]. Glycans have even been found to play a role in anhydrobiosis [6].

Promotion of folding Glycosylation promotes the folding of proteins such that more compact conformations are achieved [5]. The role of glycans in the folding process has been documented [5], with effects such as compact β turns being realised through the presence of a GlcNAc disaccharide at asparagine glycosylation sites.

Other mechanisms Glycosylation is also known to be involved in cell-cell adhesion [7] and receptor activation (such as with EGFR and TGF-β receptor [8]).

1.1.4 Disorders and disease

Genetic disorders of glycosylation have been extensively reviewed [9, 10, 11], and have been found to result in a variety of phenotypical observations. Deficiencies in glycosylation enzymes cause such phenotypes as muscle-eye brain disease, Fukuyama congenital muscular dystrophy, galactosemia and inclusion cell disease (see Table 2 in [9]). Also, the various disorders classified under the Congenital Disorder of Glycosylation (CDG) family are related to deficiencies in enzyme synthesis.

The mechanisms behind these disorders involve the disruption of various pathways as a result of the absence of a particular enzyme. By disrupting the synthesis of glycans, various structural epitopes are not exhibited, which in turn affects processes such as protein folding and lectin binding.

O-linked glycans on mucins can act as pathogen decoys, mimicking lectin receptors in human resistance to oral and mucosal infection [12]. As part of the human immune system, many lectins bind specifically to bacterial glycans, resulting in the recognition of foreign pathogens.

1.1.5 Analysis

In general, the study of glycomics involves one or more of the many techniques that can be used to glean insight into the structure, function and behaviour of glycans in a whole system [13]. Combinations of Mass Spectrometry (MS) [14] and High Performance Liquid Chromotography (HPLC) [15] are used as the primary methods of sequence analysis within glycomics. These techniques allow for qualification and quantification of glycan sequence with small amounts of sample. Two types of analysis can be performed — single structure sequencing (typically using multiple stage fragmentation MS) and glycan profiling (giving an overview of the diversity of the glycan molecules found). Each of these analysis methods is somewhat limited in the amount of data it can provide. The most information-rich analytical technique is Nuclear Magnetic Resonance spectroscopy (NMR) spectroscopy [16], which results in fully characterised structures. NMR has the drawback that large amounts of sample are required to obtain a structure.

The analysis of the data acquired through MS and HPLC requires some sophisticated bioinformatics to extract the relevant information. A number of algorithms [17, 18, 19, 20, 21] have been developed to aid in the interpretation of data (reviewed in [22]); all of them require knowledge of current structures and biosynthetic pathways.

Recently the use of various forms of glycan arrays as an analytical technique has been pursued. The two forms of glycan-related arrays are gene micro-arrays [23, 24] and glycan arrays [25]. Whereas glycan arrays allow for the parallel assessment of the lectins present, gene micro-arrays allow for expression profiling of glyco-active enzymes.

1.2 Glycobioinformatics

1.2.1 Databases

Glycobioinformatics is a nascent field with few dedicated practitioners, primarily due to the immaturity of the field as a whole. A number of groups work on the informatic analysis of glycans, each taking a distinct approach (reviewed in [26, 27, 28]). The majority of efforts in the area of glycobioinformatics have centred around the development of databases.

The development and maintenance of knowledge bases — encapsulating bodies of knowledge for a domain — are key to the practice of scientific disciplines. In molecular biology, the use of sequence databases is a standard procedure, as evidenced by genomic and proteomic disciplines using genome data and the Universal Protein resource (UniProt) resource [29] as reference databases. The free availability of reference data has spurred development in this area, opening up new avenues for research. In contrast to the genomic and proteomic databases, glycomics is missing an appropriate database to act as a reference resource.

Since the late 1980s, there have been multiple efforts to collect information on glycans and make it available through a reference database. For the majority of these efforts, database entries have been sourced from previously published data. The curation of data from literature is an expensive and time-consuming exercise, requiring at least two dedicated glycoscientist curators, database Information Technology (IT) staff, and a senior glycoscientist to oversee management. In addition, effective mechanisms to keep track of new glycan publications, and access to the journals themselves, are required.

Curation-based databases — KEGG [30], Glycosciences.de [31], BCSDB [32] and Glycominds [33] — are based upon the early Complex Carbohydrate Research Center (CCRC) Complex Carbohydrate Structural Database (CCSD) [34] (also known as CarbBank). Each database has transformed and cleaned the CarbBank data, and then made either all or a subset of the data available via the internet. Each database project then supplemented their data with new entries from additional sources. However, as each of the databases took a different approach to their design, interoperability between them is non-existent.

In addition to information from published sources, a great deal of structural and ancillary data is being generated in various glycomics labs. This data is stored in an ad-hoc format: hand-written in labbooks, stored in a proprietary electronic format, or simply sketched on a piece of paper. The lack of uniform meta-data and annotation means that over time this data will be lost and any kind of data mining cannot be performed. There is a clear need to provide methodologies for capturing this meta-data, and then providing a pathway for its eventual publication and distribution.

Currently, no data repository exists that can enumerate all the structures found in any particular glycome. However, significant interest in the area of glycomics, coupled with the large amount of data being collected in individual labs, means it is now feasible to create a comprehensive data set. Unfortunately, this data is not being captured in a uniform way, and will likely remain stored using various proprietary schemas for the foreseeable future. By creating an end-user curated database, it will be possible to create a long-lived database resource that maintains quality and reliability.

1.2.2 The need for glycan databases

It is logical to ask what the uses of glycan databases are — especially for the scientists who are being asked to fill these repositories with useful data. From a short-term perspective, there are few real perceived benefits to the end user. A more complete database may help with the implementation of various analytical tools to facilitate the analysis of data, but progress in this area is hampered more by the dearth of freely available algorithms. Glycan databases will probably provide the most utility in acting as a shortcut to finding reference information on various structures of interest. In this way, the databases act as brokers of information to the users — filtering out unnecessary data points so that the user can focus only on relevant data for further analysis. For example, Hizukuri et al. have performed data mining on the Kyoto Encyclopaedia of Genes and Glycans (KEGG) database [35] that has yielded glycan structures relevant to leukaemia [36].

This role as a filter carries through to the longer-term goals of these types of database. As the data sets approach completeness, data mining becomes more feasible. It is at this point that bioinformatics moves from its supporting role to helping to direct research. By mining the data set, it is possible to find new avenues of research which would otherwise require an inspiration to identify. Data mining essentially mimics what the human mind does when it recognises patterns in data, but on a much larger scale. By performing transformations on the data, new correlations can be identified, models tested and further experiments hypothesised.

1.2.3 Centralised curation

Databases have traditionally been maintained as centralised resources. This was often necessary because the infrastructure surrounding the database was focused on a single point of data entry, and the database servers themselves were expensive. In the last 15 years, the proliferation of commodity software and the internet have mitigated much of the expense of hosting a database.

Although equipment costs are now limited, maintenance and curation costs remain high for fully centralised databases. As the majority of data that was collected by centralised glycan databases required the review of literature and manual curation – necessitating the involvement of skilled professionals – the expense of such efforts often became prohibitive. Depending on the curation procedure, data may need to be double-checked to ensure its quality, effectively doubling the workload of entering a single entry.

A centrally curated database yields significant advantages in terms of the quality of the data, as the entire process is controlled by a central group of curators, and inconsistencies can be tightly regulated. Naturally, the curation process is unable to fix problems in the data that stem from the original publication. For example, much experimental meta-data remains unpublished due to the lack of space in journals. Although the data may be of reporting quality in the literature, it may not be detailed enough for bioinformatic analysis.

Although the results from centralised curation are good, the process is highly dependent on funding; once the money runs out, the database tends to stagnate. This is particularly evident with CarbBank and GlycoSuiteDB [37]. As the establishment of a glycome repository is a long-term project, the longevity of any database must be ensured.

1.2.4 Open curation

In contrast to centrally curated databases, open databases allow the unrestricted entry of data by interested parties. This requires no outlay for curation, but makes it difficult to maintain the quality of data. As a result of their open nature, publicly curated databases need to have measures in place to prevent bad data being inserted, either inadvertently or maliciously. These measures should attempt to ensure both syntax and context validity of the data as it enters the database.

For open curation to be successful, data must be validated without the need for centralised curation. This can be controlled either by end users, or through automatic validation procedures. To perform end-user validation, people need to be encouraged to contribute corrections to data. A study of the social methods to achieve this is beyond the scope of this thesis, but there are a number of guiding principles to making end-user validation feasible.

Ownership Users are more likely to maintain the data in a publicly curated database if they feel some ownership of its contents. Crediting data to its original source will give this person a vested interest in correcting their data, since their name and reputation is at stake.

Linus' Law Named after open-source programmer Linus Torvalds, this law states that "given enough eyeballs, all bugs are shallow". In this context, this means that with a sufficient level of peer review of records, the number of errors found in the database will tend towards zero. In addition to identifying the errors, it is also important that peer reviewers can modify the data without onerous procedural overheads. This follows the Wiki model for peer review of data.

Tooling It is critically important that suitable tooling exists to facilitate the previous two points. Appropriate tools and user interfaces remove the complexity of the data entry process, and present a simple workflow to end users. Key to the success of these tools is the timeliness of validity checking. Immediate feedback as to the context and syntax validity of entered data encourages the entry of correct data. Similarly, the benefits for the end user need to be clear — it must be obvious what the advantages of annotating data are in terms of scientific output or increased analytical capabilities.

The most technologically expedient technique for most openly curated databases is to use these manual checking methods to maintain validity. For larger databases (such as Wikipedia [38]), many of these social and procedural methods can be used. However, for lower-traffic databases, such as those found in glycomics, it becomes impractical to rely upon the general public to identify and fix errors. Instead, automatic validation procedures must be used to assess incoming data.

Ensuring syntax validity for a database can be achieved by using various automatic checking algorithms. Each record in the database must be checked against ontologies and vocabularies, which reduce database redundancy by limiting data to canonical representations. Within the field of glycomics, a good example of the practice of maintaining syntax validity is the use of well-defined sequence formats, which can be checked automatically against a well-defined grammar.

Context validity is significantly more difficult to maintain automatically in a database. Although a record may have valid syntax, the data that it represents may not be contextually correct. Context checking is a non-trivial task to automate; historically, the best approximation to this was achieved by peer review of the data. The feasibility of creating an automatic checking algorithm for context validity is examined in this thesis.

To better understand context validity, an example using the English language can be used. The sentence fragment "I can have a cheeseburger?" makes sense with respect to its syntax, following the appropriate lexical and grammatical rules. However, when placed within a context — "Two things are infinite: the universe and human stupidity; *I can have a cheeseburger?* I'm not sure about the the universe." — the sentence fragment no longer makes sense. Context validation assesses the fragment against its neighbouring data, and determines whether this new data fits with the existing data.

1.3 Summary

The study of glycans as post-translational modifications is an area that is ripe for many discoveries, especially given the links established between carbohydrates and much functionality in biological systems. The use of bioinformatics in this area has been hampered by the lack of generally available and well-curated databases. Unfortunately, obtaining such databases requires a significant investment of money in the curation process. The open curation model can provide similar results, but extra technology needs to be in place to compensate for the lack of human review of database entries. It is as part of a database review procedure that the algorithms in this thesis have been developed — to allow for an automated check of the context validity of structures.

Within this thesis, Chapter 2 introduces the biosynthesis of glycans and the biological basis for context validation. The algorithms for data verification and validation, as well as details about the developed software, are found in Chapter 3. Chapter 4 introduces the main results from the thesis and provides an estimate as to the size of the glycome. The work done to establish the distributed component of the EUROCarbDB database comprises Chapter 5. Chapter 6 summarises the full thesis, and provides an outlook for future developments in this field.

Chapter 2

Biology of structural verification

The quality of data in a database can be evaluated by examining two key inherent properties of the data. The first critical property of the data is syntax validity. Lexical errors in sequences must be minimised, with facilities provided to ensure that the entered data matches what the depositor intended to enter. Syntax validity can be ensured using lexical parsers, the description of which is beyond the scope of this thesis.

The second critical property of the data is context validity. Despite syntax validity, there is no guarantee that the data is correct with respect to — or in the context of — other content in the database. Although the responsibility of entering contextually correct data lies largely with the depositor, automatic checks can be performed to evaluate each new sequence to see how well it fits into its particular context. For new data that is not closely associated with existing data, the onus is on the depositor to justify its addition to the database.

To check context, new sequences must be evaluated against the currently known sequences for the biological source in question. The natural metric for how well one sequence is associated with another is sequence similarity. It is not possible to use a naive alphabet-based sequence similarity metric against a generalised glycome-wide sequence set since there is no guarantee that the known sequences themselves are similar. In fact, the use of any method to check syntactic similarity will always fail in the presence of protein sequence family heterogeneity. Basic biosynthetic pathways are generally conserved across sub-systems within a glycome; with this in mind, a sequence similarity algorithm that compares sequences against the current biosynthetic pathway knowledge base has been developed, and forms part of this thesis.

It is important to clarify that any glycome discussed here is not a complete reference glycome. Since the methods of regulation of the biosynthetic pathways leading to glycosylation have not been fully elucidated, no single set of sequences can be said to represent perfect knowledge about the glycosylation occurring in a particular biological system. We know neither the end sequences nor the mechanisms of synthesis. In a practical sense, the glycome represents only the set of sequences that have been deemed interesting enough by researchers to both characterise and enter into a database.

The current glycan biosynthetic knowledge base is codified within pathway rules, which determine the order of action of a series of enzymes on various substrates to gain end products. A simplified model of enzyme reactions — in which each enzyme reaction can be modelled as a linear application of substrate, donor and product — is used for this study.

Biosynthetic pathway information in the glycan space has traditionally found limited use in bioinformatics. Informally, it supplements incomplete data when performing sequence assignment. General explorations of the glycan space have also been undertaken [39], but no commonly accepted number that can act as a measure of the diversity of human glycans has been established. A number of rough estimates have been made — from 500 structures [40], to tens of thousands [41] to 10-10^4 times larger than the proteome, which is conservatively estimated at 470,000 glycan structures [10].

2.1 Biosynthesis of glycans

Unlike proteins, the biosynthesis of glycans does not follow a template-based model [1]. The process of glycosylation is complex, with many factors affecting the end glycosylation products. Glycan biosynthesis occurs primarily in the Golgi apparatus, where the glyco-active enzymes have their effect. There are several families of glyco-active enzymes: the glycosyltransferases (EC 2.4.x.x), the glycosidases (EC 3.2.x.x), sulfotransferases (EC 2.8.x.x) and phosphotransferases (EC 2.7.x.x). The spatial location, competition between, and concentration of these molecules (together with substrate and donor concentration) are all factors in determining the end glycosylation products for any system.

Glyco-active enzymes (also known as glycogenes) are regulated on the transcriptional level. A subset of the genes is expressed across all tissues, while the rest (notably those responsible for non reducing end terminal diversification) are expressed on a tissue-by-tissue basis [41, 42]. This difference in expression accounts for some of the glycans' structural diversity. Tissue-specific expression also correlates with functional groupings of tissues, suggesting that the synthesis of certain glycan structural features is relevant to the function of the tissue [43, 39].

Glycosyltransferases and glycosidases are both membrane-bound proteins that build and prune glycan structures, respectively. Glycosyltransferases catalyse the transfer of a monosaccharide unit, typically from a donor molecule to a defined substrate acceptor. The donor molecules are usually sugar nucleotides, such as UDP-glucose, UDP-galactose or UDP-GlcNAc.

Glycosidases degrade glycans through the hydrolysis of a glycosidic bond, and have varied specificities. Glycosidases can be specific to both anomeric or monosaccharide conformations. Degradation of glycans allows structures to be remodelled, exposing different epitopes to change the function of the glycosylation, as well as modifying the action of any further glyco-active enzymes by altering the target substrates.

Glyco-active enzymes are highly variable in the reactions that they catalyse. It is not sufficient to assume that a glycosyltransferase will catalyse a single reaction with a single substrate, nor does the absence of a glycotransferase directly imply the absence of the linkage in the synthesised structure. Linkages are redundantly coded in different transferases, and a transferase can also catalyse different linkages. Substrate specificity and donor availability both play parts in deciding which reactions are catalysed by a particular glyco-active enzyme. Although the multitude of enzymes catalysing a linkage can imply redundancy along biosynthetic pathways, the redundant genes may not be expressed in the same tissue, resulting in non-redundant expression of linkage catalysis capability.

Gene expression A subset of genes was found to be differentially expressed in different tissues in a PCR study of expression patterns [42]. Of 67 genes, only 35 were expressed in all tissues. Genes that were only partially expressed across all tissues were indicative of regulation at the transcriptional level. Of particular interest, all the genes responsible for the catalysis of the GlcNAc($\beta1 \rightarrow 6$) $\Rightarrow$ Gal and Fuc($\alpha1 \rightarrow 3/4$) $\Rightarrow$ GlcNAc disaccharides are not expressed in all tissues, suggesting some regulation of branching on the transcriptional level. Most of the disialylation genes (ST8SIA) are expressed on a per-tissue basis, while the tissue distribution of the core Glc and GalNAc transferases suggests that each gene may code for different substrate specificities, regulated on the transcriptional level.

After protein synthesis in the endoplasmic reticulum (ER), proteins are modified as they pass through the Golgi apparatus. The Golgi apparatus is arranged as a series of cisternae (see Figure 2.1), in a stack-like formation, with inter-cisternae movement achieved through vesicular transportation.

There are two theories as to the arrangement of the glycosylation machinery within the Golgi apparatus, according to the two current models of Golgi maturation [44]. One maturation model anchors glyco-enzymes in the Golgi, with the glycosylation substrate being transported between cisternae via vesicular transport. The other model localises the proteins to be modified in the cisternae, transporting the glyco-enzymes in a retrograde fashion back up the Golgi stack. Regardless of which model is correct, both imply a compartmentalisation of the enzymes, resulting in the sequential application of enzymes to proteins that are to be modified. Complexes can be formed to further cement the ordering of the application of enzymes [45, 46, 47], which ensures their sequential application through their proximity both to each other and to the target glycosylation substrate.

Beyond transcriptional regulation, regulatory mechanisms are required to account for the structural diversity of glycans. Within a single tissue — and indeed, even on a single protein — glycosylation is not necessarily uniform. Despite this, there are core structural motifs shared amongst these structures, suggesting that this micro-heterogeneity is controlled from within the

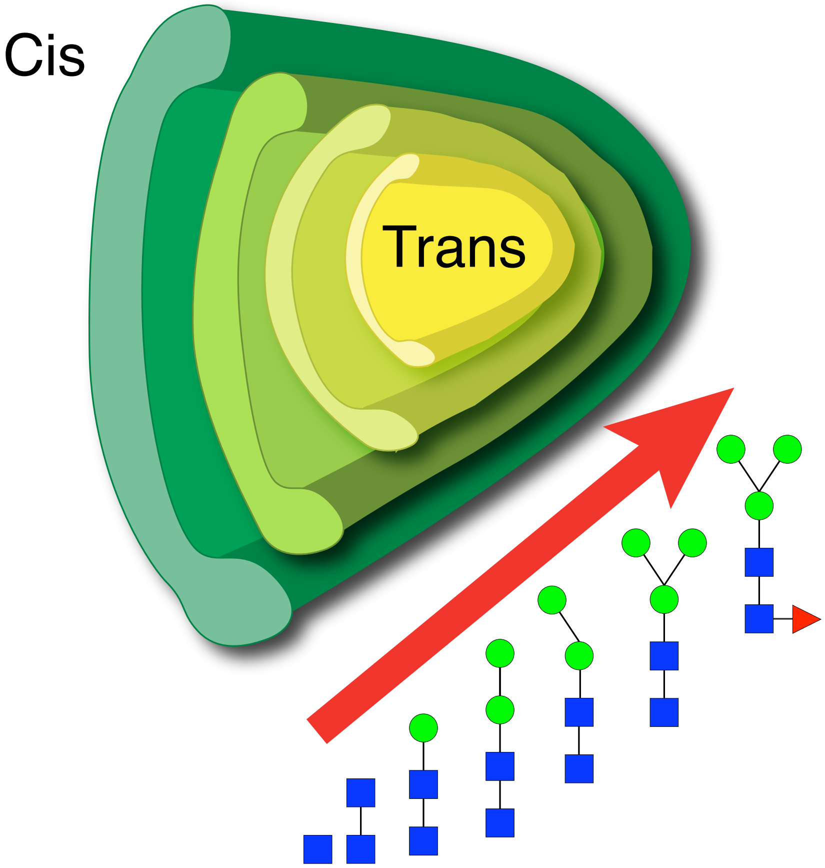


Figure 2.1: The Golgi apparatus is a series of cisternae, arranged from cis to trans. Protein cargo moves from cis to trans, and various enzymatic treatments are applied. As glycans pass through the Golgi apparatus, enzymes are applied sequentially, resulting in a gradient of increasing complexity of glycans.

Golgi apparatus. Alongside the compartmentalisation of the various enzymes, donor concentration and the availability of precursor substrates all play a part in the production of the diverse set of structures.

Donor concentration plays a significant role in the regulation of glycosylation. The interactions between donor concentration, substrate availability and enzyme efficiency can lead to sophisticated glycosylation behaviours. One such example is the effect of N-Acetyl-D-glucosamine (GlcNAc) concentration on the branching behaviour of N-glycans. Branching on N-glycans displays an ultrasensitive relation to hexosamine concentration [8], an increase in the concentration of GlcNAc residues results in an increase in branching, thanks to the differing specificity and substrate competition behaviour exhibited by the relevant enzymes. As a result of this ultrasensitivity, glycosylation profiles can be modified relatively rapidly during the cell life-cycle, providing a finer-grained control mechanism for the regulation of glycosylation. Donor concentration can in turn be regulated in earlier pathways such as the fructose or mannose pathway (KEGG pathway 51).

In addition to these regulatory mechanisms, other factors have an impact on the degree and type of glycosylation. Protein conformation plays a role in determining the specificity of enzymes (such as GALNT5 [48]) to particular substrates, by means of sequence recognition, patch recognition, and changes in the accessibility of the oligosaccharide substrate. Since it is possible that the glycosylation on distinct and structurally different proteins can occur in parallel, the recognition of protein [49] or lipid substrate as part of the acceptor substrate offers a possible explanation for the heterogeneity in glycosylation profiles across proteins being glycosylated at the same time. In addition, since glycosylation can affect the conformation of proteins [50], it is plausible to posit that any glycosylation event may influence subsequent glycosylation. Since glycosyltransferases compete for substrates, it is possible that other proteins being glycosylated in the vicinity may affect the glycosylation of a protein by preferentially using up donor molecules, leaving it with only lower-affinity glycosylation reactions.

Defects in the glycosylation machinery commonly lead to a corresponding profile exhibiting incomplete glycosylation [51]. That glycosylation is exhibited at all suggests that there is no glycosylation quality control mechanism in the Golgi apparatus, implying that the partially glycosylated substrates will remain available in the Golgi pipeline even if essential glycosylation is not completed. Genetic defects in the glycosylation machinery are known as CDG, and have been reviewed extensively [10].

The specificity of enzymes varies — from the previously mentioned conformation-specific GALNT5, to enzymes that recognise a disaccharide structural feature (such as B4GALNT1/2, synthesising the SD$^{\alpha}$ epitope), or less specific enzymes, such as glycosidases.

Based upon this regulatory information, the glycosylation machinery can be generalised into two regulatory categories. The first combines transcriptional regulation, compartmentalisation, and enzyme specificity to provide coarse control over glycosylation. The second uses finer-grained control mechanisms such as donor concentration and compartmentalised substrate competition to allow for temporal modification of the glycosylation profiles — especially for terminal structures.

Functional studies of glycosylation have been undertaken through the engineering of various knock-out strains of mice [52] which exhibit both differing phenotype and changed glycosylation profiles. Interestingly, the changes in glycosylation profiles between knock-out mice have either been drastic, or simply changed the quantities of the exhibited structures. Drastic profile changes have been associated with mice that knock out early stage glycosyltransferases in the pathway, whereas knock-outs of the later-stage transferases (such as the decorative FUT enzymes) cause more subtle changes. This lends credence to the concept of glycosylation being tuned at later stages, with this secondary means of regulation being supplemental to the transcriptional level.

Depending on the nature of the change in expression levels for a particular glycosyltransferase, different effects can be seen. Underexpression of an enzyme will most likely result in a cascade effect, in which other enzymes requiring substrates catalysed by this particular enzyme cannot function. This results in a whole family of structures not being synthesised. Overexpression, or even introducing new enzymes into a system, may not necessarily change the glycosylation profile. The absence of donor residues, or the limited amount of available substrate, may conspire to reduce the effect of changes in the enzyme's expression level. Generally speaking, there are too many factors involved to be able to gauge the effect of overexpression.

2.2 Bioinformatic analyses of pathway information

Given the inherent complexity of the glycosylation machinery, and the subtlety of the operation, this area is ripe for rigorous bioinformatic analysis. In particular, it is practical to take a pluralistic systems biology approach to analysis. The study of glycomics data has thus far not been generally applied in the field of systems biology — primarily because there is no well-defined algorithm for predicting glycosylation from the varied set of inputs that regulate it, and as such, we do not know all the forms of glycosylation which may exist. Similarly, there is no clear understanding of the functions of glycans on their own; further study is needed to elucidate this in a generalised way.

All bioinformatic analyses regarding biosynthetic pathway information has thus far focused on using this data to predict N-linked glycosylation profiles. N-linked glycosylation has been the most well-studied of glycosylation types, and all structural databases reflect this bias. The wealth of data surrounding N-glycans makes this family of structures a logical target for building models.

The most sophisticated modelling method for glycosylation is the mathematical model approach [53, 54], which models the cisternae in the Golgi as a series of chemical reactors with a constant flow rate, using differential equations to determine the success of a particular reaction. These models have been used to predict the whole glycosylation profile in response to modifications to donor and enzyme concentration. The modelling has been undertaken by groups looking to produce uniform glycosylation profiles by searching for the optimal enzyme and donor concentrations. The utility of this method is dependent on having good data about the enzymes being modelled, as it involves *in silico* fine-tuning of the glycosylation machinery to produce the desired structures.

In combination with gene expression data, biosynthesis data can be used to attempt to predict glycan structure profiles [39]. By considering the co-occurrence of disaccharides in structures, donor-substrate reactions can be clustered together. The clusters of reactions reveal classes of structures as well as biosynthetic pathways. By finding the structures with the closest number of common donor-substrate pairs, the best predicted structures can be suggested. A limitation of this methodology is that the fine-tuning of the glycan structures is not taken into account — the suggested structures will be representative of the larger-scale regulation of glycosylation.

Estimates of actual glycan diversity are quite difficult to make. Upper bounds on the total numbers of structures can be provided, through basic combinatorial methods, but the actual number is further limited by certain biochemical factors. Currently, the optimal way to estimate the size of the glycome is to extrapolate from existing databases, which are an approximation of what is at the very least biosynthetically possible.

There are two approaches to the classification of glyco-active enzymes — functional and sequence. Functional classifications come in the form of EC numbers — with 295 glycosyltransferases and 189 glycosidases identified. Because a single enzyme can have many different functions, it is possible for more than one EC number to be assigned to a particular gene. Sequence classification, exemplified in the CaZY [55, 56] classification system, examines the polypeptide sequence similarity of enzymes. This classification does not necessarily have a functional relation, and is of limited use unless studying the actual enzymes.

2.3 Bias in existing structural databases

Current knowledge about structural diversity can be obtained by examining entries within existing databases. Over the past 20 years, a number of structures have been deposited into the databases in order to facilitate the creation of these global knowledge bases. The vast majority of the data has been created *ad hoc*; there have been no concerted efforts to provide whole glycomes for any one system. The diversity of structural sources inherent in the databases has a number of implications. The most important is that it is difficult to draw any conclusions through analysis of the databases as to the completeness of any one set of structures.

Because a number of techniques have been used to elucidate the structures found in databases, it is unknown whether the structures are representative of the complete glycome or are simply the most easily qualifiable structures. A large amount of general biosynthetic pathway knowledge has been used to assign sequences, so the entries in the databases may also be biased towards structures which fit into known biosynthetic pathways.

This possible bias in the sequence databases makes it difficult to perform any meaningful bioinformatics in the realm of glycomics. Any conclusions drawn must be treated with great skepticism — if biosynthetic pathways are in agreement with sequence data, this may provide insight less into the actual extent of the glycome, and more into the use of pathway information in sequence assignment. Steps can be taken to mitigate this, such as extensively verifying literature and comparing independently validated curated databases.

2.4 Human system as a model system

There are a number of estimates as to the size of the human glycome. A lower estimate of the size of the human glycome is 500 [40], which is less than the number of human sequences in most databases. Higher estimates put the number of glycans at 10-10^4 times larger than the proteome — at a conservative estimate 47,000 – 470,000 glycan structures [10], tens of thousands of structures [41], or thousands of times more complicated than the genome.

The human system has been chosen here as the target system for data analysis due to the congruence of the availability of data in terms of genes, expression, sequence and structure. This wealth of information is a result of this system being well studied. Although the completeness of glycosylation profiles for various sub-systems varies across the whole system, we can still use this data to draw general conclusions about the nature of glycosylation in humans, with certain caveats. A number of glycan classes are expressed — from the sub-types of N-glycan, to O-glycans and proteoglycans. This diversity of expression also makes this a rewarding system to analyse. Currently, the best approximation for a human glycome will come from using a subset of the GlycomeDB [57] to obtain the structures. To verify the data, a set of independent structures are obtained from GlycoSuiteDB [37], which is current to 2005.

2.5 Summary

The biosynthesis of glycans in the ER and Golgi is a complex process, involving many subsystems, and dependent mechanisms. The system can be better understood and analysed by using biosynthetic pathway data, which can be considered representative of all collected data in databases, and possibly even the glycome as a whole. Pathway data may therefore be used as a surrogate for the complete database in order to perform contextual validation. This must be done cautiously, however, as inherent biases may affect the efficacy of the validation; pathway data must always be verified. Assuming that biosynthetic information can be encapsulated in pathway data, it also becomes feasible to estimate the size of the glycome from this information.

Chapter 3

≈

Methods

To be able to validate sequences against a surrogate for contextual information — i.e. pathway data, we need to establish the integrity of the pathway data to see whether it can be used to judge the novelty of structures entered into the database. This involves verifying that the enzymes can code for the synthesis of all structural features. Subsequently, the pathway data needs to be evaluated for its ability to support the synthesis of these features. Any differences in the support of enzyme and pathway data of the complete feature set points to new structural features which are not covered by current knowledge bases, or possible errors in the structures that need to be resolved. Once the pathways have been verified, a metric to convey the differences between the candidate structure and the current knowledge base can be obtained, and the size of the glycome estimated.

This complete analysis can be broken down into two steps — data collection and verification. Data collection and verification must take place independently for the enzyme data and pathway data.

3.1 Enzyme data

3.1.1 Data collection

Data collection for enzyme data was undertaken using a manual curation process. Initially, a list of enzymes was obtained from the KEGG Glycan [30] resource. Each gene name was associated with a particular donor and acceptor substrate known to catalyse a reaction and was generalised as: Donor($anomer\ linkage$) $\Rightarrow$ Substrate. Since different identifiers were being used to identify each of the genes from the various resources, the names were normalised to use a generalised identifier scheme based upon the Entrez GeneID. As the rate of enzyme discovery is low, it was deemed unnecessary to automate this initial data collection step. The data collection yielded a set of disaccharide reactions with broad substrate specificities — often recognising only a single residue on the target substrate. To supplement this data, the GGDB [58] was sourced to obtain more reaction substrate specificities. Once again, a manual curation process was used to obtain this data, capturing as much substrate information as possible. It was originally envisaged that further substrate specificity information could be obtained through a literature survey. To this end, a web-based curation tool was developed to allow for the rapid entry of data, whilst also allowing for simple searching of existing data.

For the initial data collection phase, a tabular approach was used, based upon a spreadsheet application. This obviated the need for further software to be developed and maintained while the data being collected was still undefined. Once the data set reached a satisfactory size (approximately 50 entries, enough to understand the nature of the data), a more rigorous format for the storage of data was decided upon, and a database system designed to hold the data. Sequences for donors and substrates were initially encoded in the International Union of Pure and Applied Chemistry (IUPAC) condensed format [59] (GlycoSuiteDB variation) with no support for repeating units or uncertainty in sequence elements. Sequence encoding was moved to a variation of GlycoCT [60], to permit comparison with various databases of structures. Data were designed to be exported and imported from the database using a simple markup to identify the fields in the exchange document. For this purpose, a XML file with versioning for reproducibility was used.

3.1.2 Data verification

	Group 1	Group 2	Group 3	Group 4
Genetic information	Y	N	Y	N
Database verification	Y	Y	N	N

Table 3.1: Reaction classification. Reactions can be classified into one of four groups, depending on the availability of genetic information and the status of database verification.

To verify the enzyme data, reactions were placed into four groups (Table 3.1). Reactions that are backed up with genetic information (Groups 1 and 3) are defined to be correct, whereas reactions for which evidence exists only in databases (Groups 2 and 4) are deemed to be potentially inaccurate. The potentially inaccurate reactions are the most important to examine to evaluate the source of the inaccuracies.

For a reaction to be validated using the database alone, the product of the reaction needs to have been observed in real structures. A reaction can be observed if the resultant disaccharide which the reaction catalyses is found in a structure. For example — if the Gal(β1→4)GlcNAc linkage is found in the database, this provides evidence for the correctness of a Gal(β1→4) $\Rightarrow$ GlcNAc reaction. All structures need to be searched for the presence of the disaccharide, so all disaccharides need to be generated for the structures.

To obtain the disaccharides from the database, a simple breadth-first tree traversal — in which the child residues for a particular residue are enumerated — is performed on each structure in the

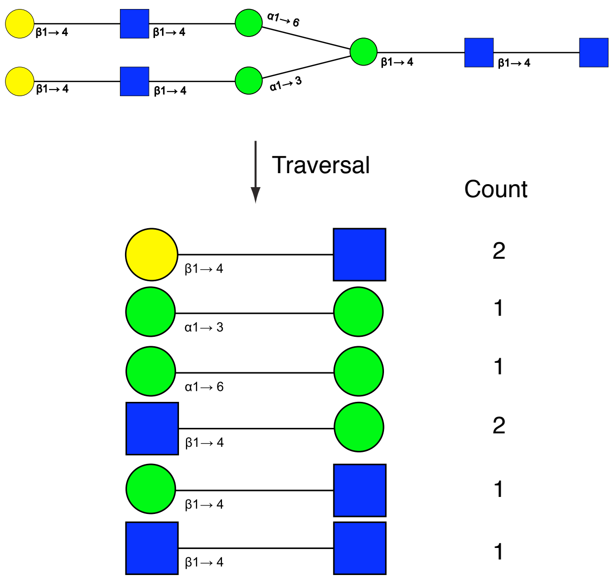
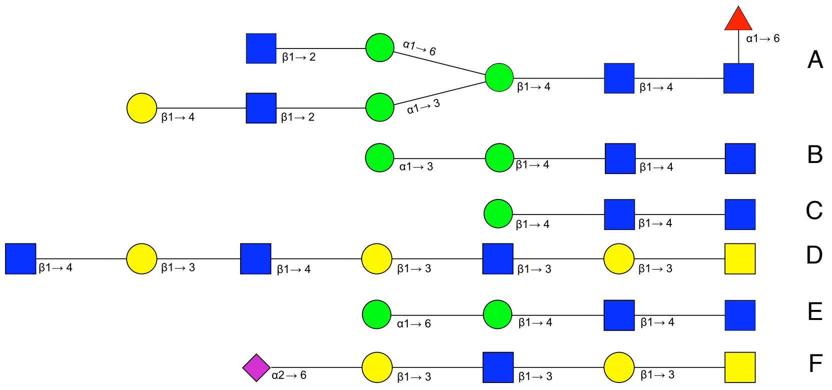


Figure 3.1: Generation of disaccharides from a structure. A traversal of each of the parent-child residue pairs is performed, resulting in a set of unique disaccharides and the frequency of their occurrence in the structure.

database (see Figure 3.1). The identifier for the structure is also assigned to the disaccharide to allow the data to be traced back to its originating structure. This step allows for the generation and counts of all disaccharides for a structure. The disaccharide lists for each structure are then combined with the disaccharides for all other structures. Multiple occurrences of a particular disaccharide in the one structure are only counted once, since there is only one structure that serves as evidence for that particular disaccharide (see Figure 3.2). All the disaccharides for a particular set of structures can be listed in this way.

Since the database itself is unreliable, we need to perform filtering on the generated data to ensure that the set presented contains only disaccharides known to occur in a minimum number of structures. For the purposes of this analysis, the filtering is set to a disaccharide appearing in a minimum of six structures. After this first filter, further filtering occurs whereby disaccharides from structures that have been entered into the database incorrectly are removed. This is a manual curation step, and is largely dependent on further meta-data on the original structure being available for examination. Secondary filtering was performed for data coming from the Glycosciences.de database and Carbbank [34], as the reference information was readily available for examination. For certain disaccharides, structures were examined based upon their original papers. From here, errors in sequence were looked for, and if found the originating structure was deemed invalid. Errors found in sequences ranged from incorrect transcription of data from the



	A	B	C	D	E	F	Total
🟢 α1→3 🟢	1	1					2
🟢 α1→6 🟢					1		1
◆ α2→6 🟡						1	1
🟡 β1→3 🟦				2		1	2
🟡 β1→4 🟦	1						1
🟦 β1→3 🟡				1		1	2
🟢 β1→4 🟦	1	1	1		1		4
🟦 β1→2 🟢	2						1
🟦 β1→4 🟦	1	1	1		1		4
🟡 β1→3 🟨				1		1	2
🔺 α1→6 🟦	1						1

Figure 3.2: Collating the set of disaccharides from a set of structures. The disaccharides are generated for structures A-F, with the individual frequency for each of the disaccharides found listed. The totals are derived by counting the non-zero frequencies for each disaccharide across all structures. A frequency greater than 1 for a structure is not counted more than once since the structure only provides a single piece of evidence that the disaccharide exists in nature.

paper, incorrect encoding of the sequence within the paper to incorrect assignment of sequence as a result of flawed analysis. The incorrect assignment of sequence within the paper was determined by examining the techniques used to elucidate sequence, and judging if there is sufficient data to make the claims as to the existence of the structural feature. At a bare minimum, for a disaccharide deemed to be new, a technique which could unambiguously identify the structural feature was required to be used as evidential material. This evidence could include lectin specificity, glycosidase specificity, the presence of diagnostic ions within MS data, or the full characterisation of the structure using NMR. Often, structures were assigned to have a particular sequence with insufficient evidence to support the assertion. Appendix A lists all the structures from Glycosciences.de that were determined to have errors associated with them.

Following these filtering steps, a manual comparison step was performed to determine which of the disaccharides and reactions were actually observed, and which were database artefacts. The results can be found in Chapter 4.

3.2 Pathway data

3.2.1 Data collection

Pathway name
N-Glycan biosynthesis
High-mannose type N-glycan biosynthesis
N-Glycan degradation
O-Glycan biosynthesis
Chondroitin sulfate biosynthesis
Heparan sulfate biosynthesis
Keratan sulfate biosynthesis
Glycosaminoglycan degradation
Lipopolysaccharide biosynthesis
Peptidoglycan biosynthesis
GPI-anchor biosynthesis
Glycosphingolipid biosynthesis — Lactoseries
Glycosphingolipid biosynthesis — Neo-lactoseries
Glycosphingolipid biosynthesis — Globoseries
Glycosphingolipid biosynthesis — Ganglioseries

Table 3.2: Names of pathways used in the analysis of pathway data

Collection of pathway information is significantly less complex as the KEGG pathway resource is the primary source of information for this data. A File Transfer Protocol (FTP) download allows the data for all the different pathway classes to be obtained. The biosynthetic pathways available from KEGG are listed in Table 3.2.

Informatically, a representation of a pathway can be realised by considering it as a state machine. A state machine is a graph in which the nodes represent states, and the edges represent transitions between states. For the purposes of pathways, the substrates and end products of reactions are nodes, while the enzymes involved are edges. Within the downloaded data files from KEGG, the end products and substrates are stored as KEGG structure identifiers. To translate the KEGG identifiers to actual sequences, GlycomeDB [57] was used to obtain the sequences in GlycoCT [60] format. Since each reaction and substrate pair appears only once in a pathway, it was sufficient to use the reaction definition from the enzyme database, and annotate it with a pathway identifier to indicate that it is part of a pathway. Starting from the smallest structures in a pathway, we can recreate the pathway state machine by looking for all reactions with the pathway annotation, matching end products with substrates.

3.2.2 Data verification

Pathway verification is an involved process. Essentially we wish to verify whether the structures found at steps along the pathway can be found in nature, and that all structures in nature can have their biosynthesis explained by a pathway. Simplistically, we can examine the structures in each state of the machine, and verify that there are structures which cover that part of the state machine. This method fails for coverage of the entire pathway, since there are intermediate structures which may not have been characterised experimentally, although structures later in the pathway may exist. This method also only gives a simple characterisation of the data deficiency — it can only characterise the presence or absence of the data. Ideally, an analysis would better characterise the pathway data with respect to known structures to identify any potential new biosynthetic pathways.

One convenient way to calculate the difference between the known structures and what is defined in the pathway is to calculate the composite structure map — with counts for the number of times a particular structural feature is seen — and then to subtract the largest pathway from this larger composite structure (Figure 3.3). Although simple, this methodology does not differentiate between the extensions that can take place along different points along the pathway, and does not allow individual pathway extensions to be identified.

To allow for individual extensions to be accounted for, one can take the largest structure from a pathway, and calculate intersections of this structure with all other structures (Figure 3.4). Any structure which contains the largest structure as a sub-structure can be considered to have an extension to the defined pathway of some sort. By storing the sequence of each of the extensions, one can list the extensions to a maximal pathway individually. This method is flawed as it does not

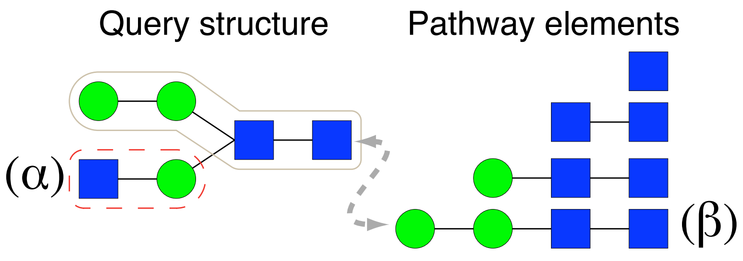


Figure 3.3: Resolving a single structure using pathway elements. The query structure is compared with the largest tetrasaccharide from the pathway structures (β). The two residues GlcNAc and Man are identified as extensions (α) on the largest pathway element, while the matched tetrasaccharide is marked in grey. This basic step identifies a single extension on the largest element of a pathway, and needs to be expanded so that it can identify multiple extensions along the pathway.

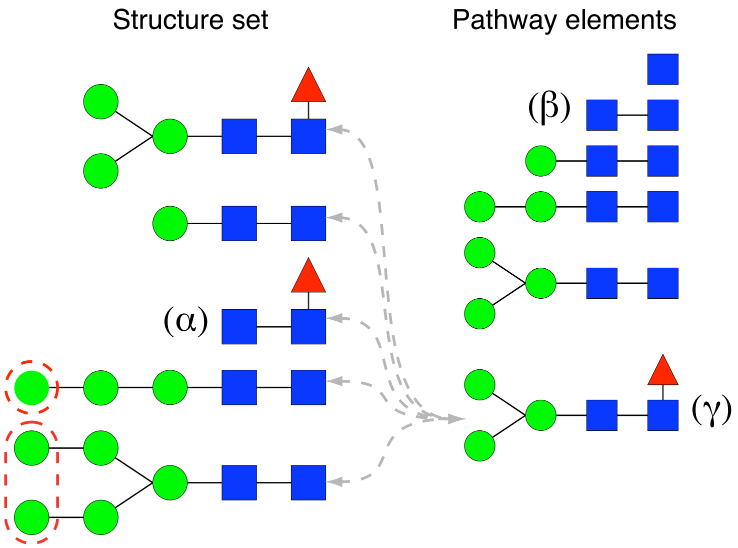


Figure 3.4: The largest pathway element (γ) is compared with all structures in the structure set. Any extensions found using this algorithm are marked in red. The fucosylated di-GlcNAc structure (α) has not been found to have an extension, even though the fucosylation occurs on a substrate which has not been identified as fucosylated on the pathway elements (β).

account for extensions occurring exclusively at earlier points (such as branching decision points) along the pathway. A method that discovers these individual extensions occurring along any point along the pathway is needed.

To discover the full multitude of extensions, we need to take each structure from the database, find the largest pathway structure that is smaller than that structure, and then collect the extensions to the pathway structure — including the points at which the extensions are attached. This process is repeated for all of the structures in the particular subset of data that is being examined. After this, the additions at each unique attachment point are combined into one structure, giving the full set of extensions at the location in the pathway where the attachment point is exposed. The path from this point to the root of the glycan is called the root path. By merging only the root paths and pathway extensions together, we can create a composite structure that contains only the extensions along the complete pathway. This algorithm is described in Figure 3.5. Since each of these activities is a non-destructive merging activity, it is possible to keep a tally of how many structures each of these extensions appears in. This can be used later to filter out any structural features which result from errors in the data, by removing low-abundance sequence noise.

Following generation of the new composite extension maps, each extension that has not been covered by a pathway needs to be resolved, and examined to see if it represents a potential new pathway. The automatically converted pathway data from KEGG is not complete — certain extensions involving repeat structures are not encoded into the pathway data, and some elements may not have been translated correctly from the KEGG data into the local pathway data. At the end of this process, it should be possible to determine the data completeness and the extensions required so that it covers a greater number of structures.

When evaluating the coverage of the pathways, it was convenient to consider fucosylation, sialylation and sulfation events to occur outside the pathway mechanism, as these structural features occur at many points along the pathway. They are essentially modifications upon the backbone of the structure, whose distribution needs to be examined separately.

3.2.3 Mathematical operators for glycans

Glycans are naturally represented as graph-like constructs (see Section 3.4). For the structural analysis of glycans, a set of basic graph-like operations is of great utility. The operations that are most useful are those related to set theory. Union and intersection form the basis of the addition and subtraction operations upon glycans respectively.

Unions of glycans can only be performed when they have at least one residue at a particular position in common. For example, this could mean that both sugars share a common root residue. The union operation is actually a special case of a traversal operation, where each node is visited, and any children that are not part of the first addend are cloned and attached to the first addend, as shown in Figure 3.6. A node cloning operation is a deep cloning, and all descendants of the

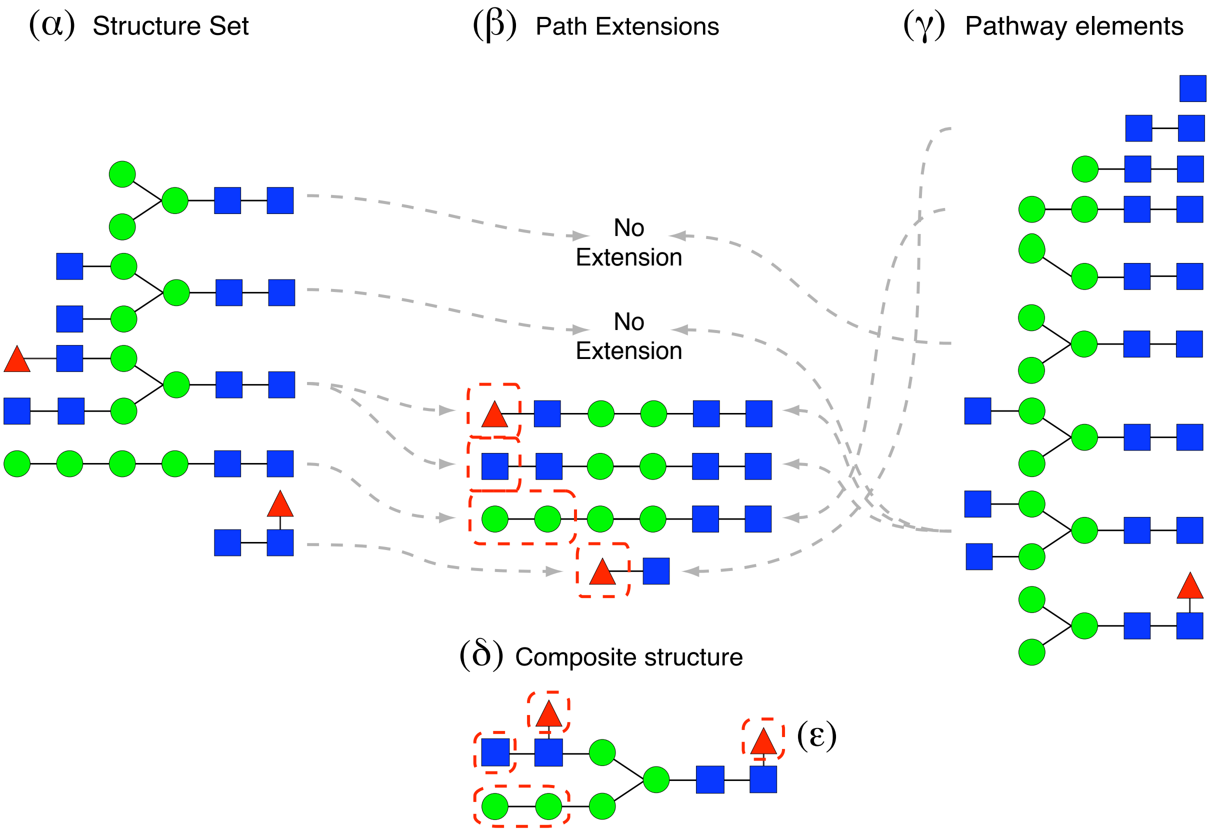


Figure 3.5: The final algorithm for detection of extensions along the pathway. Each structure from the structure set (α) is compared with the largest pathway element from the pathway elements (γ) that are contained within the structure. From this, the extensions are identified, as well as the path back to the root, which is used to identify the location where the extension is applied. A union is then performed on the path extension set (β) so that a single extension labelled composite structure is generated. This composite structure (δ) identifies new extensions along the pathway, as well as structural features which are synthesised at points earlier in the pathway (ϵ).

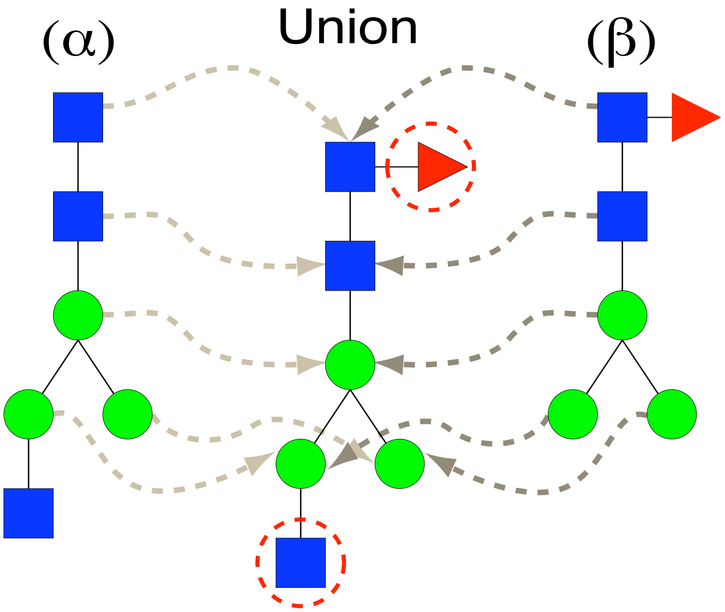


Figure 3.6: The union of structures (α) and (β). A parallel traversal is performed on structures (α) and (β), where the children of each node are compared. Any children not found in both sugars are added to the resultant sugar (marked out in red) so that the resultant structure contains all structural features found in structures (α) and (β).

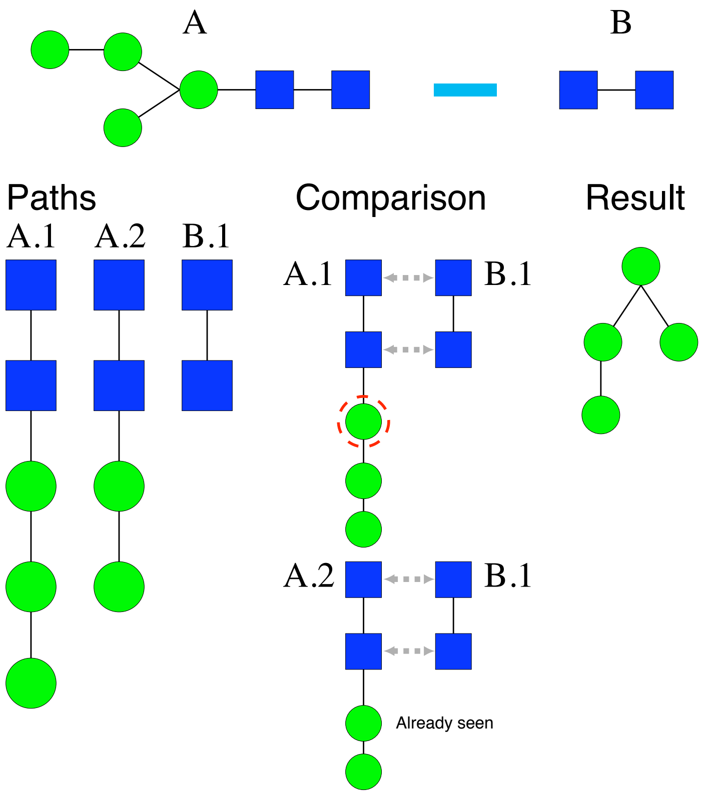


Figure 3.7: Subtraction of a sugar. Paths A.1, A.2 and B.1 are generated from structures A and B by performing a traversal from leaf to root. Paths A.1 and B.1; and A.2 and B.1 are traversed in parallel to compare nodes. The residue marked in red is deemed to be unique to the first structure, and the sub-tree rooted at the residue is taken as a result. On the second traversal, this same residue is found, but is not a new result since it is already part of the result set.

node are added. In all cases, the sugar is defined as an edge-labelled tree structure — where each individual edge is unique. Circular structures are not abundant enough in human data to warrant adding support for these types of structure in the analysis. By performing a union across two graphs, the labels on the edges may not necessarily be unique (although in combination with the node they attach to, they form a unique tuple), resulting in the need to modify basic tree traversal algorithms to handle this multiplicity. In order to simplify the process of performing the union, glycans are restricted to those forming only acyclic graphs, with a common label on both the root nodes.

The algorithm for calculating the intersection of two glycans is essential. A greedy algorithm is used, growing the maximal common sub-tree from the root of the first operand. Each of the paths from the leaves to the root of the structure is reversed, and is followed in parallel down the other operand. Any nodes that are not members of the path are excluded from the intersected structure. The algorithm for subtraction (Figure 3.7) is an extension of the intersection algorithm, where sub-trees originating from the excluded nodes are returned as the non-common nodes.

3.2.4 Glycosidase modelling

Glycosidases are not explicitly modelled within the system. Since each of the glycosyltransferases catalyses the addition of only a single residue onto a substrate, it is sufficient to model the activity of glycosidases by considering that the action of a particular glycosidase is equivalent to the inaction of the complementary glycosyltransferase. Also, given that the main pathway in which glycosidases are active is within the N-linked pathway, where mannose trimming takes place, it was deemed unnecessary to model the action of glycosidases. Within pathways, all substrates that are the end product of the action of a glycosidase are already modelled within the system. Since the algorithms are designed to find extensions upon the pathway, this provides us with enough information to find any possible extensions from this point onwards.

3.3 Validation of sequences

The actual algorithm for the validation of sequences is a specialisation of the algorithm to validate pathway data against the database. Only a single structure (instead of the entire database) is checked against the set of pathways. Any parts of the structure that are not covered by the pathway are deemed to be one of several defined branch extensions derived from the results section. If any residues cannot be accounted for through the pathway or branch extension, then the structure is deemed as potentially invalid for the context.

3.4 Software models and implementation

There are two object models used to realise the algorithms. The first is a simple data storage model, used to maintain the integrity of the enzyme data. The second model is a set of programmatic entities which can be used as a toolkit to perform these analyses.

3.4.1 Model for data storage

The storage model used for enzymes is simple, with weak relationships defined between the entities. The entities can be broken down into Enzymes, Genes, Reactions and References. The UML diagram (Figure 3.8) explains the relationship between these entities, and the fields that they each have.

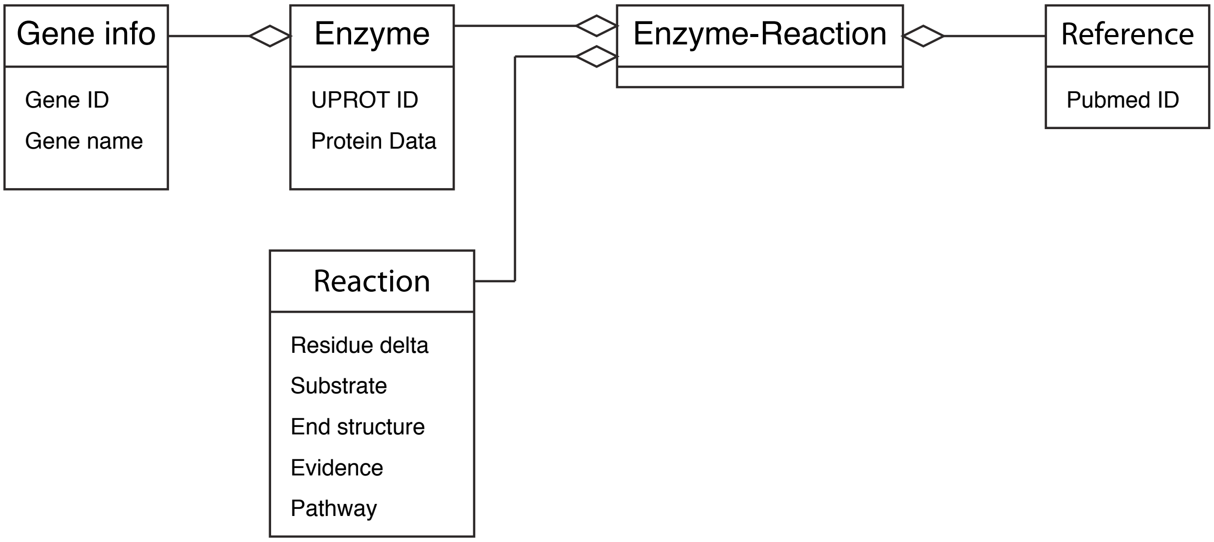


Figure 3.8: UML class diagram for the storage model of enzymatic data. Data types are not shown for fields within the model. The model was designed to allow for a generic collection of data concerning enzyme specificity on the protein level. In a practical sense, the model was used more to link the gene information to the reactions that each gene may catalyse.

3.4.2 Software model for analysis

The analysis model is based on the manipulation of a series of classes that model the sequence of a single glycan. As with any toolkit for analysis, the flexibility of the tools used is critical. Priorities in the development of the software were maintaining the modular functionality of the toolkit, as well as procedures to ensure reproducibility of results.

3.4.3 Software implementation for analysis

To perform the data analysis, custom software was written. A conscious decision was made not to support the modelling of all the structural features found on glycans, as this adds complexity to the software, which can only be used in limited situations. Instead, the software toolkit was designed to be flexible enough to perform a variety of structural analyses. In addition, the toolkit was designed so that simple scripts can control the execution of analyses, and so that all results are easily reproducible, even when faced with changing data and algorithms.

Multiple namespaces

Since this project was in part a data integration project, and data could come from a number of sources, residues on the sugars were given identifiers which were de-coupled from their name in any particular sequence format. So this means that the name for a residue can be obtained in GlycoCT, Glyde [61] or IUPAC condensed at runtime. This was of particular benefit for writing the test suites for the software — all test cases could be written in IUPAC condensed format, and the same program logic would be used as if the sequences were written using GlycoCT. In addition, tests could be made on residue names across different namespaces, according to what best suited the program being run. For example, tree traversal algorithms remained the same while a single namespace switch would result in the algorithms working on residue super-classes.

Reporting modes

A sophisticated rendering system was developed to allow for the generation of interactive reports on the data. Rich reporting was key to the project, as one of the main challenges with interpretation of resultant data is the large volume of it. Intermediate composite structure results had more than 100 residues within structures, each with several annotations on the residues. Existing rendering methods only provided static rendering of sequences — any further annotation on sequences could not be rendered as part of the report.

The central component of the report-rendering engine is a basic structural layout engine and various plug-in rendering engines. The basic structural layout engine was used to provide a simple layout of the structures, so that all structural features could be observed. It was not a priority to produce images that matched with the established layout of sugars, so drawing features like fucosylation as a stub residue was not implemented. The most important rendering engine developed was the Scalable Vector Graphics (SVG) rendering engine, which produced vector graphic rendering of the sugars. By using XML-based rendering, it was also possible to carry through the annotation of structures into the report document itself through the class attribute of image elements. This document could then be used to generate static images — in Portable Network Graphic (PNG) format for example — or could be embedded in an eXtensible HTML (XHTML) page to act as a starting point for further exploring the data set in interactive reports.

The ability to explore the data set, and the rich reporting used during the analysis, greatly aided in the interpretation of data. Any manual annotations on the data were saved, so that all reports could be manually regenerated.

Unit testing

Due to the modular nature of software development, it was prudent to use a comprehensive set of test suites to ensure that the core algorithms worked as expected, and that no changes in the algorithms would change the results in any way — known as a regression. Each change set in the core objects was associated with a test case, which allowed the added functionality to be tested, as well as any regressions in functionality which may have been introduced by the addition of this functionality.

Language choice

Ruby [62] is a powerful language, due to high level programming constructs and meta-programming features. The language itself allows for a semi-mathematical expression of algorithms, and is well suited to the job of list processing and filtering — which is the basis of most of the algorithms used in this project. Based upon the strengths of the language, Ruby was chosen as the language in which to implement the software.

Web application

Part of the original plan for this project was to supplement the original enzyme data with manually curated data from literature, in order to build a comprehensive database of all enzymatic information. Although this part of the project was not used, a web application was built using Ruby on Rails to assist in the maintenance of the data. Simple operations for the addition, deletion and modification of entities were automatically generated, and then more complex reporting pages were created.

In addition, some small applications were created to allow for exploration of data, and the easy deposition of any new information. One tool looks at the enzyme coverage map for a particular structure, another generates the full set of theoretical structures from a set of enzymes, and another was developed to serve as a test for a completely graphical method of drawing structures. All these tools remain in an early stage of development, as they were designed primarily to aid in the interpretation of the data.

Since the application logic is de-coupled from the presentation, it was simple to add pure XML and plain text reports on the data in the database, which also allowed for more complex functionality to be built up by mixing and matching existing functionality.

3.5 Summary

To develop a technique for automatic sequence validation with respect to context, a number of verification methods needed to be established. For the biological system in question (Homo sapiens in this case), enzyme and pathway data was obtained. Enzyme data can be curated from various databases available on the internet, while pathway data is available from KEGG. Both the enzyme and pathway data then needs to be verified. By comparing the enzyme and pathway data with structures in the database, it is possible to gauge how well this data represents the sequences within the database. The algorithms to perform this can be realised in Ruby, or any other suitable high-level language. Once it is established that the pathway data is representative of the data in the database, it can be used to validate new sequences in the context of the biological system, and also to estimate the size of the glycome.

CHAPTER 4

RESULTS

TO VALIDATE STRUCTURES using biosynthetic pathway data, data was collected and verified from various sources. The verification of enzymes and pathways yielded a number of results. Specifically, a number of key findings with respect to the correctness of enzyme data and pathway coverage were established. Secondary investigations involving substrate analysis were also performed. After the verification process, it was possible to develop an algorithm to perform automatic context validation, and an estimate of the size of the human glycome was provided.

4.1 Enzyme verification

The enzymes curated from the data sources are shown in Section A.1 (found in Appendix A). There are 136 distinct genes identified as being responsible for the action of glycosyltransferases, which in turn have been associated with a total of 303 different reactions. Of these only 60 substrate/donor residue and linkage groups are unique. This is a larger number of enzymes and reactions than the KEGG analysis [39], which had only 98 enzymes and 42 reactions.

The enzyme reactions were compared with the 255 different disaccharides found in the human subset of GlycomeDB. Table 4.1 lists the donor, acceptor and linkage combinations, highlighting the entries where either no genetic information has been collected for a linkage, or where genetic information has been found, but no structures with that particular disaccharide have been found. A filter was applied to the disaccharide data so that only the disaccharides that have been found in at least six independent structures are displayed. Applying this filter results in the total number of disaccharides being reduced to 53. Linkages in green are supported by enzyme data, linkages in black have no supporting structural data, linkages in grey are common structural errors, whereas linkages in blue do not have any supporting enzymatic data. The literature used to support the presence of these new linkages is listed alongside the disaccharides in Table 4.1.

Interestingly, the Gal(α1→3) ⇒ GalNAc reaction, falling below the tolerance level for inclusion in the table, has been identified as a unique structural feature in carcinoma [74] and on mucins [75] [76]. The Gal(β1→3) ⇒ Glc reaction motif also falls below the tolerance level for inclusion in the

	Gal	GalNAc	Glc	GlcN	GlcNAc	GlcA	Man	NeuAc	NeuGc	Xyl	Fuc
Gal	$\alpha1\to3$ $\alpha1\to4$ $\beta1\to3$ $\beta1\to6$ [63] [64]	$\beta1\to3$	$\beta1\to4$		$\beta1\to3$ $\beta1\to4$					$\beta1\to4$	
GalNAc	$\alpha1\to3$ $\beta1\to3$ $\beta1\to4$	$\alpha1\to3$ $\beta1\to3$ [65]			$\beta1\to4$ $\beta1\to3$	$\beta1\to4$					
Glc			$\alpha1\to2$ $\alpha1\to3$ $\alpha1\to4$ [66] $\alpha1\to6$ [66]		$\beta1\to4$		$\alpha1\to3$				
GlcNAc	$\alpha1\to4$ $\beta1\to3$ $\beta1\to4$ [67] $\beta1\to6$	$\beta1\to3$ $\beta1\to6$			$\beta1\to4$	$\beta1\to4$	$\alpha1\to2$ $\beta1\to2$ $\beta1\to4$ $\beta1\to6$				
GlcA	$\beta1\to3$	$\beta1\to3$			$\beta1\to3$ $\beta1\to4$						
Man				$\alpha1\to4$	$\alpha1\to4$ $\beta1\to4$		$\alpha1\to2$ $\alpha1\to3$ $\alpha1\to6$ $\beta1\to3$ $\beta1\to6$				
NeuAc	$\alpha2\to3$ $\alpha2\to6$	$\alpha2\to3$ [68] $\alpha2\to6$			$\alpha2\to6$ [69]			$\alpha2\to8$			
NeuGc	$\alpha2\to3$										
Xyl										$\beta1\to4$	
Fuc	$\alpha1\to2$ $\alpha1\to3$ [71] [72] $\alpha1\to6$ [73]		$\alpha1\to3$ [70]		$\alpha1\to3$ $\alpha1\to4$ $\alpha1\to6$						

Table 4.1: Human disaccharides derived from the database compared with collected enzymes. Green linkages have both enzyme and database support, blue have only database support and black have sonly enzyme support. Grey linkages have been found to be database errors. Disaccharides occurring fewer than six times are filtered out.

table, but has been found to exist in urine [77], although the structure may have originated in a non-human system [78]. Similarly, the Gal(β1→6) ⇒ Gal reaction has been detected in urine [63] [64]. Since these structures may not be synthesised by the endogenous glycosylation machinery, it may be prudent to exclude urinary glycans from any kind of future whole glycome analyses. In practice, this may be hard to achieve, as this localisation data may not be stored in the database.

On the other hand, the GalNAc(β1→3) ⇒ GlcNAc reaction has not been found in any structures. The enzyme catalysing this reaction (B3GALNT2) has been characterised [79], but structures containing this feature have not yet been found. EXT1 and EXT2 encode the GlcA(β1→4) ⇒ GlcNAc reaction found in heparan sulfate, but there is no disaccharide data to support this. This is due to the encoding of this linkage within repeat structures, which have been excluded from the main body of the disaccharide analysis. The Man(α1→4) ⇒ GlcN reaction (encoded by PIGM) is present only in GPI anchor biosynthesis — the structures of which are not found in the structural databases.

A subset of reactions is supported both by structures and by enzymatic information, but the structures in which these motifs are found occur too seldom to pass the structure count tolerance filter. The resulting disaccharides form transferases that take fucose as the acceptor substrate — B3GALT1 (catalysing a Glc(β1→3) ⇒ Fuc), LFNG and RFNG (responsible for the GlcNAc(β1→3)Fuc linkage) are all under-represented in the database. The EXTL linkage GlcNAc(α1→4)GlcA) also appears infrequently, even though it is found in the heparan sulfate biosynthesis pathway. This is to be expected, given the lack of evidence for the linkages catalysed by EXT1 and EXT2.

The number of filtered human disaccharides (53) compares well with the number of reactions collected from glycosyltransferase databases (60). Although a filter was used for this data, there are still 11 reactions found in human structures that have not been associated with either enzymatic or pathway information. There are many more reactions that did not pass the structure count filter, and without checking the references for all those reactions, it is not possible to pass judgement as to whether these structural features are real, or whether they are artefacts. In general however, the similar number of reactions that has been curated and generated indicates that the size of the curated reaction set is of the right order of magnitude.

A summary of mammalian glycosyltransferases [11] can also be used to compare the completeness of the enzyme data collected. Comparison of the reaction matrix from this paper with the results here indicats a number of discrepancies (Table 4.2).

4.2 Substrate elucidation

It became clear during the data collection phase that the substrate specificities for the enzymes were not well established. This is largely due to the appropriate data not being present. To synthesise the Sd$^{\alpha}$ epitope, the GalNAc(β1→4) ⇒ Gal reaction requires the presence of the NeuAc(α2→3)Gal substrate [80], and so the NeuAc residue should always be a sibling of the GalNAc residue — that

Found exclusively in the Database	**Found exclusively in Summary**
Fuc(α1→3) ⇒ Gal	GalNAc(α1→6) ⇒ GalNAc
Fuc(α1→6) ⇒ Gal	GlcNAc(α1→6) ⇒ GlcNAc
Gal(β1→6) ⇒ Gal	GlcNAc(α1→4) ⇒ GlcA
Glc(α1→4) ⇒ Glc	GlcA(β1→4) ⇒ Gal
Glc(α1→6) ⇒ Glc	Man(α1→4) ⇒ GlcNAc
GlcNAc(α1→4) ⇒ Gal	Xyl(α1→3) ⇒ Glc
GlcNAc(β1→4) ⇒ Gal	Xyl(α1→3) ⇒ Xyl
GlcNAc(β1→3) ⇒ GalNAc	
GlcNAc(β1→4) ⇒ Man	
GlcNAc(β1→6) ⇒ Man	
NeuAc(α2→3) ⇒ GalNAc	
NeuAc(α2→6) ⇒ GlcNAc	
Xyl(β1→3) ⇒ Xyl	

Table 4.2: A comparison of lists of reactions found exclusively in collected data versus a review paper.

is the NeuAc and GalNAc residues both share the same parent. To verify this, and to identify any substrates which require the presence of a particular sibling, the sibling residues were counted. A number of these siblings were identified, and can be considered as branching points on the structures. If a particular residue is almost always seen next to another residue, it can be assumed that the pathway has an influence on the attached residue, one residue is required for the other to have an action, or a combination of the two. If the relationship is not reflective, then the action of this particular transferase is optional — and is equivalent to a branching step.

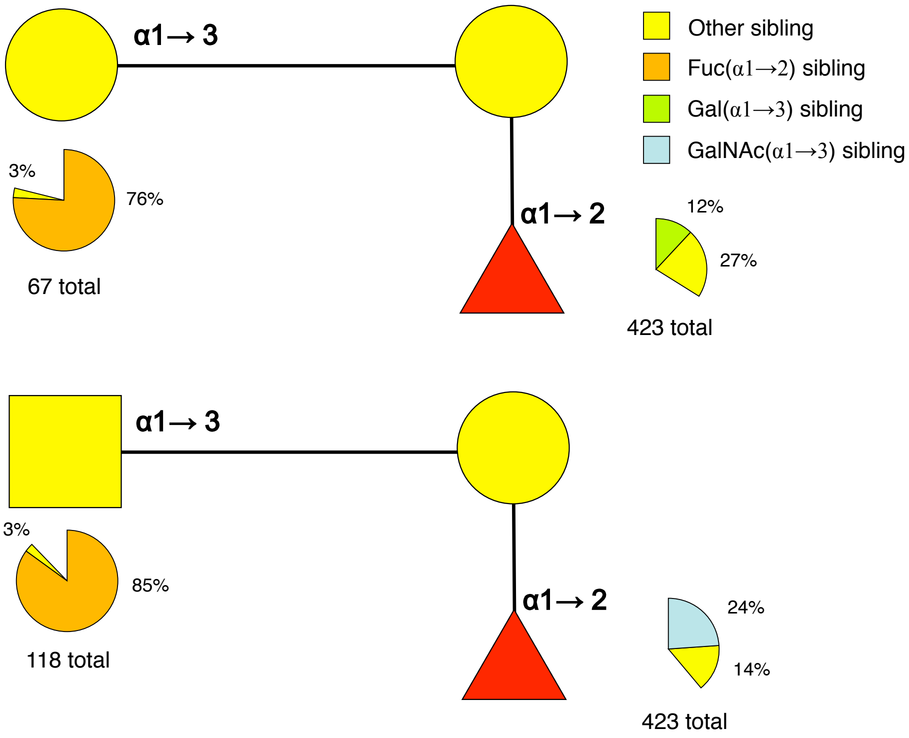


Figure 4.1: Fuc(α1→2)Gal substrate. The pie charts show the number of siblings that are found for each residue of each type.

Fuc(α1→2)Gal substrate See Figure 4.1. The Fuc linkage primarily occurs with no siblings (61%), but whenever it is found at a position where branching occurs, it is as a sibling of either Gal or GalNAc. The GalNAc linkage has a sibling 88% of the time, which is the Fuc residue in 85 of the cases. Similarly, Gal has a sibling 79% of the time, which is primarily (76%) a Fuc residue. Within pathways, the Gal(α1→3) ⇒ Gal and GalNAc(α1→3) ⇒ Gal reactions occur sequentially after the Fuc(α1→2) ⇒ Gal reactions. This data suggests that the reactions catalysing the GalNAc and Gal linkage both preferentially act upon the Fuc(α1→2)Gal substrate, or that the galactosyltransferase and GalNAc transferase may form a complex with the fucosyltransferase.

NeuAc(α2→3)Gal substrate See Figure 4.2. In 75% of cases where the GalNAc(β1→4)Gal disaccharide is seen in structures, a sibling residue to GalNAc can be found. A NeuAc residue on the α2→3 linkage is the predominant (91% of total) residue to form a sibling to GalNAc. The NeuAc(α2→3)Gal disaccharide occurs primarily at a non-branching linkage position, having no sibling 93% of the time. Ninety-six percent of the NeuAc(α2→3)Gal linkage siblings are GalNAc on a β1→4 linkage. Given these numbers, it is impossible to draw the conclusion that the

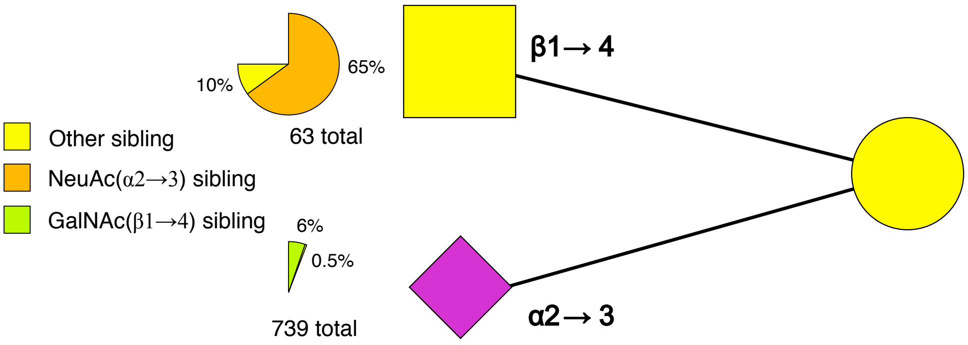


Figure 4.2: NeuAc(α2→3)Gal substrate. The pie charts show the number of siblings that are found for each residue of each type.

GalNAc(β1→4) transferase absolutely requires the NeuAc(α2→3)Gal substrate. However, pathway data supports this conclusion, as the GalNAc transferase has its action at two separate points along the pathway — one which uses sialylated substrate, and the other which uses the simple Gal substrate residue.

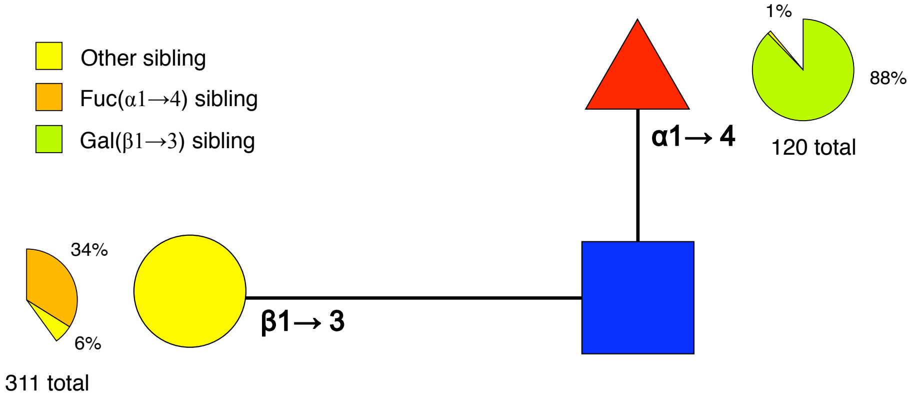


Figure 4.3: Gal(β1→3)GlcNAc substrate. The pie charts show the number of siblings that are found for each residue of each type.

Gal(β1→3)GlcNAc substrate See Figure 4.3. The Gal → GlcNAc disaccharide primarily occurs in the absence of siblings (60%), with the remaining 40% found in concert with Fuc(α1→4)GlcNAc (34%) and NeuAc(α2→6)GlcNAc (6%). The Fuc linkage is found as a sibling of other linkages 90% of the time, of which the Gal → GlcNAc linkage is a sibling in 99% of cases. Additionally, the Fuc linkage is a sibling with NeuAc(α2→6) three times — a very small number of cases. The data suggests that the Fuc linkage shows a substrate preference for the Gal(β1→3)GlcNAc substrate.

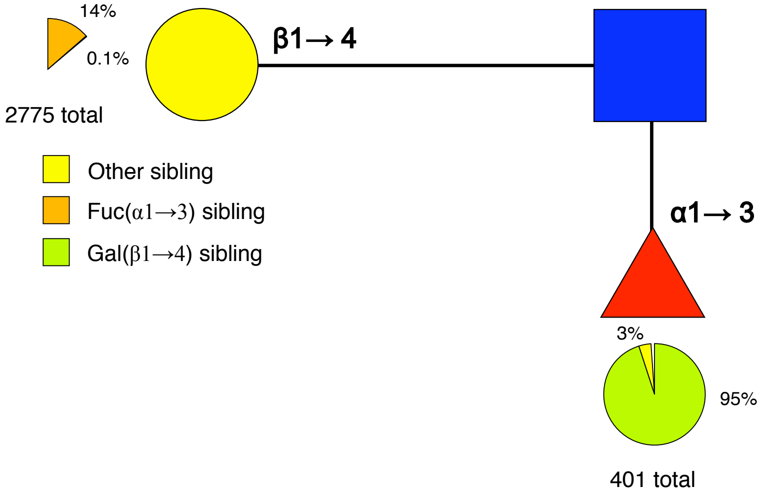


Figure 4.4: Gal(β1→4)GlcNAc substrate. The pie charts show the number of siblings that are found for each residue of each type.

Gal(β1→4)GlcNAc substrate See Figure 4.4. The Gal(β1→4)GlcNAc disaccharide is also primarily a non-branching linkage (86% of cases). Of the 14% of cases where this disaccharide marks a branch point, the Fuc residue is the primary sibling residue. The Fuc residue is linked on more than one linkage, but the most predominant is the α1→3 linkage (99%). The Fuc residue is a sibling with other residues 98% of the time, predominantly Gal on a β1→4 linkage (95%). This data also suggests that the Fuc linkage shows a substrate preference for the Gal(β1→4)GlcNAc linkage.

GalNAc(β1→4)GlcNAc substrate This linkage is found to have no branches in most of the cases where the structure has been seen, except for a minimal number of Fuc(α1→3)GlcNAc linkages, suggesting that the GalNAc(β1→4)GlcNAc linkage is a non-preferential substrate for the Fuc ⇒ GlcNAc transferase.

Gal(β1→3)GalNAc substrate See Figure 4.5. The GlcNAc(β1→6)GalNAc disaccharide occurs predominantly in a branching position, with siblings in 92% of observations. Of the siblings, the Gal residue on a β1→3 linkage is the most prevalent, occurring in 81% of cases. The Gal(β1→3)GalNAc disaccharide is mixed between having siblings and being a non-branching feature (branching occurring in 46% of structures). For the branching cases, the siblings are mixed between NeuAc(α2→6) (20% of all siblings) and GlcNAc (78% of all siblings) and other residues for the remaining 2%. This suggests that the GlcNAc is a branching feature on the Gal(β1→3)GalNAc, which can optionally

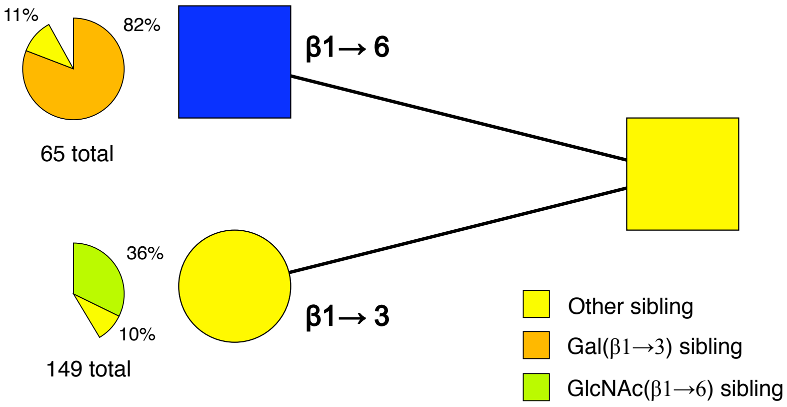


Figure 4.5: Gal(β1→3)GalNAc substrate. The pie charts show the number of siblings that are found for each residue of each type.

occur when the NeuAc has not capped the structure (the GlcNAc(β1→6)GalNAc disaccharide has not been observed with any NeuAc siblings).

GlcNAc(β1→?)Man substrate See Figure 4.6. Each of the GlcNAc → Man disaccharides has a different number of siblings attached to it, depending on the linkage position of the GlcNAc on Man. The 6-linked disaccharide has a sibling in 96% of occurrences, the 4-linked in 94% and the 2-linked in 40%. The 6-linked is found as a sibling to the 2-linked GlcNAc in 98% of cases where there is a sibling. The 4-linked is found as a sibling to the Man(α1→3/6) disaccharides, where it acts as a bisecting GlcNAc linkage in 40% of the sibling cases, and to 2-linked GlcNAc in 77% of cases. The 2-linked GlcNAc is a sibling of 6-linked GlcNAc in 43% of the observations, and 4-linked GlcNAc in 55%. There are a very small number of cases (5) where 4- and 6-linked GlcNAc linkages are siblings to each other, suggesting that the 2-linked linkage is optionally branched by the 4-linked and 6-linked GlcNAc. The 6-linked GlcNAc exhibits a strong preference towards the 2-linked GlcNAc as a sibling.

GlcNAc(β1→4)GlcNAc substrate See Figure 4.7. The GlcNAc(β1→4)GlcNAc structure is optionally fucosylated on the first GlcNAc, and the sibling data reflects this. The Fuc(α1→6) disaccharide has a sibling in 99% of observations, which are almost exclusively (98%) GlcNAc residues on a β1→4 linkage. This disaccharide itself has a sibling in 51% of cases, of which 99% are the Fuc residue on a α1→6 linkage. This data strongly suggests that the Fuc(α1→6) has an extremely strong preference for the GlcNAc(β1→4)GlcNAc substrate.

GlcNAc(β1→3)Gal substrate See Figure 4.8. GlcNAc(β1→6)Gal has a sibling in 82% of cases. Of these, it is sibling with GlcNAc(β1→3) in 98% of observations. GlcNAc(β1→3) has a sibling in

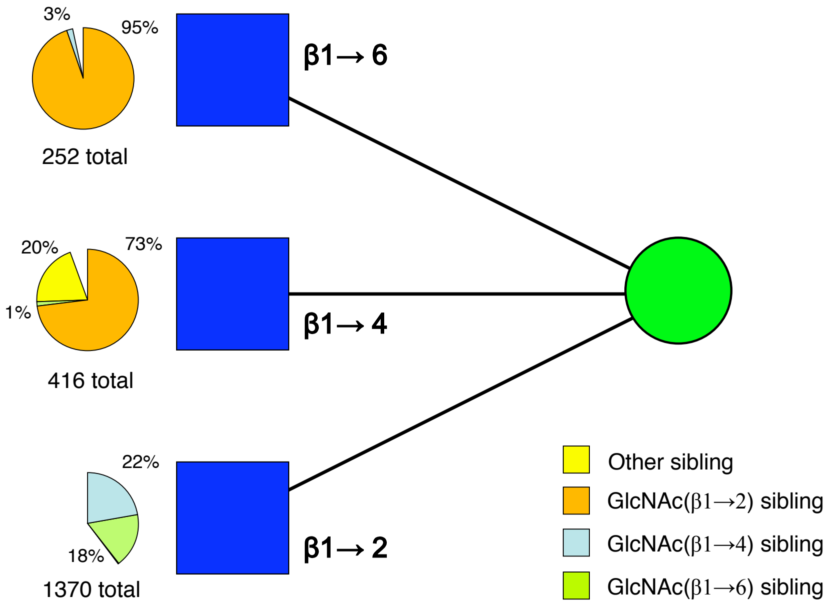


Figure 4.6: GlcNAc(β1→?)Man substrate. The pie charts show the number of siblings that are found for each residue of each type.

only 20% of cases, but of these cases it is sibling with GlcNAc(β1→6) in 89% of observations. This suggests that the GlcNAc(β1→6) linkage preferentially binds to the GlcNAc(β1→3)Gal site, and matches with its behaviour as a polylactosamine branching site.

NeuAc(α2→6)Gal and Man(β1→4)GlcNAc substrates Both these linkages occur primarily at non-branching points, with siblings in 1.7% and 0.1% of cases respectively. The NeuAc linkage is a capping linkage, whereas the key role that the Man linkage plays in N-glycan synthesis and the position within the pathway may explain why no other additions are seen at this position.

Man(α1→2)Man substrate This linkage is unique in that it has absolutely no observed siblings, suggesting that for the Man(α1→2)Man disaccharide, this substrate is only really suited towards further extension by further Man(α1→2)Man linkages.

A number of disaccharide structural features have been identified through examining the occurrences of siblings, and possible substrate preferences have been shown. The action by which these preferences are realised could be through a compound of enzymes, or through mechanisms such as substrate competition. It is not possible from the data to draw exact conclusions about the nature of the reactions catalysing these linkages, but the observations should provide starting points for further experimental investigation.

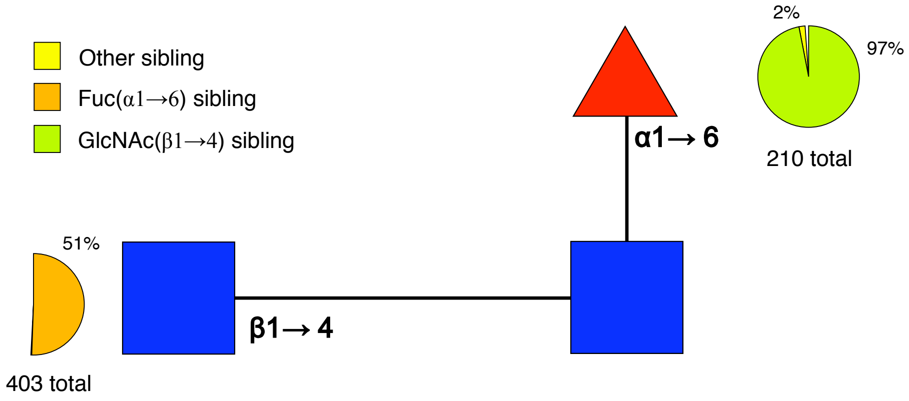


Figure 4.7: GlcNAc(β1→4)GlcNAc substrate. The pie charts show the number of siblings that are found for each residue of each type.

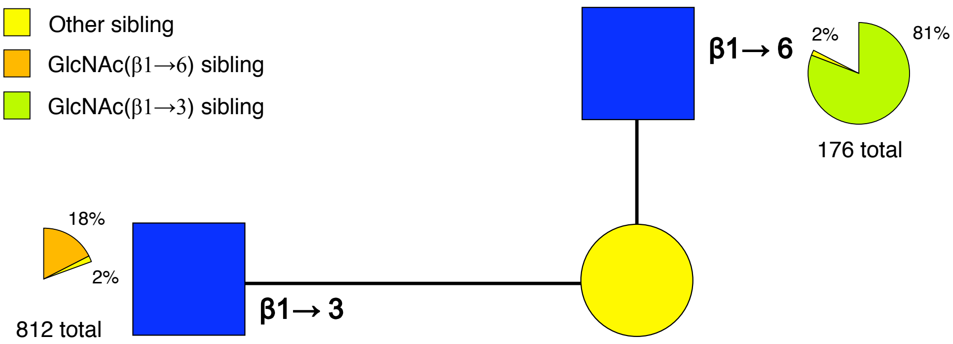


Figure 4.8: GlcNAc(β1→3)Gal substrate. The pie charts show the number of siblings that are found for each residue of each type.

4.3 Pathway verification

Composite structures comprising all the structures found in the pathways can be found in Section A.3. Running the analysis against structures from GlycomeDB and GlycoSuiteDB, a number of coverage comparison maps were obtained for each of the pathways. In order to verify the existence of the structures, two of the most populated pathways were used to compare the structures found in GlycoSuiteDB and GlycomeDB (Figure 4.9 and Figure 4.11). As shown in the composite map, GlycoSuiteDB and GlycomeDB structures have a great deal of structural features in common. It is important to note that the structural features found in GlycomeDB are a superset of the structural features found in GlycoSuiteDB. In addition, the number of occurrences of structural features were compared. There were no significant differences in the distribution of structural features — that is, apart from a few branches, there are no large differences between the distribution of structural topology families between the databases. Since GlycoSuiteDB and GlycomeDB are independently curated databases, we can draw the conclusion that if a structure is contained in both databases, it is far more likely to have a valid syntax due to the redundancy in data collection.

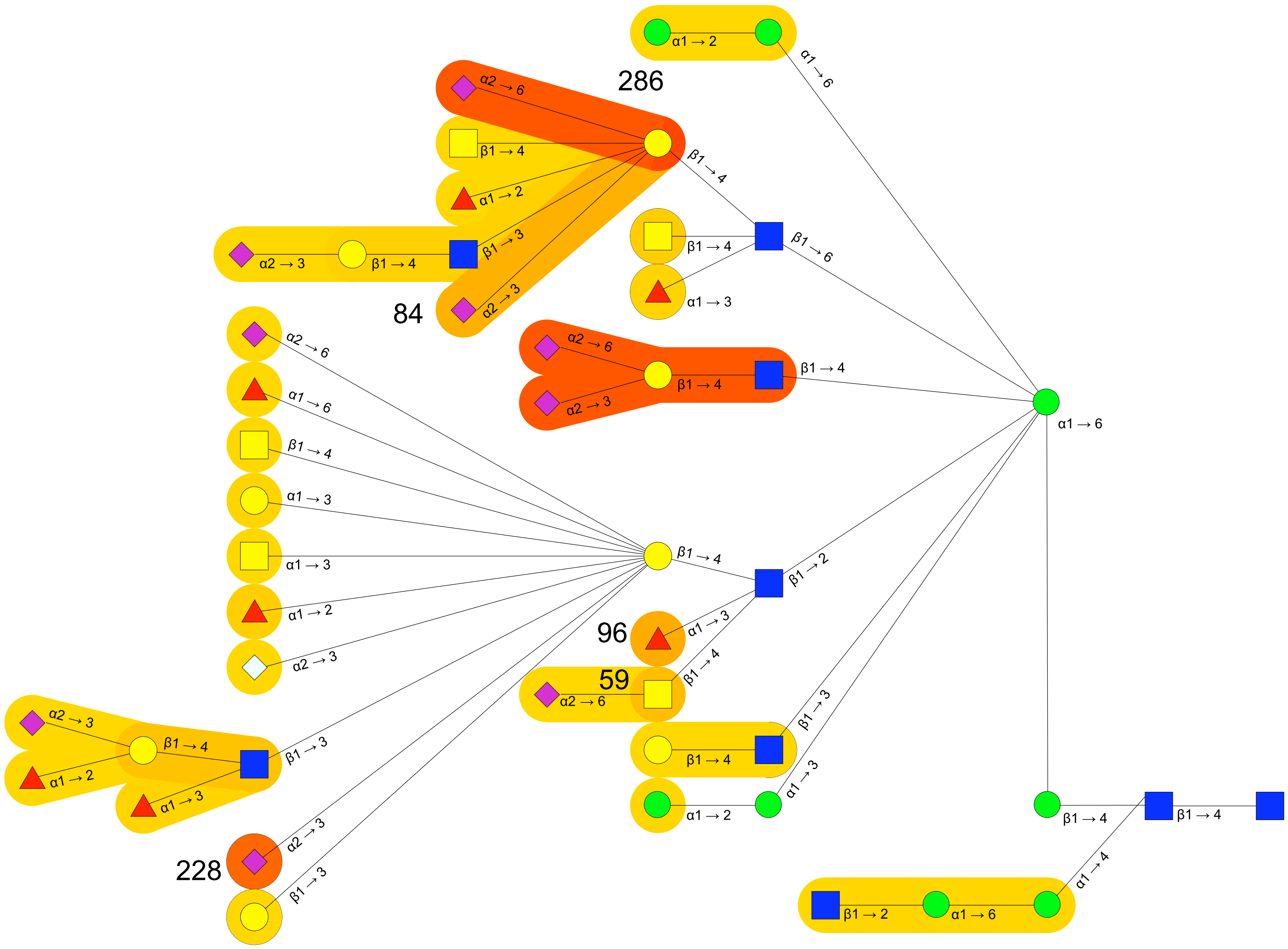


Figure 4.9: Comparison of composite structures from GlycoSuiteDB and GlycomeDB along the N-linked pathway. Differences in frequency of features are mapped by heat — red features being more commonly found in GlycomeDB. Residues without any heat information are those found in the pathway that are common to both databases. There are no structural features unique to GlycoSuiteDB.

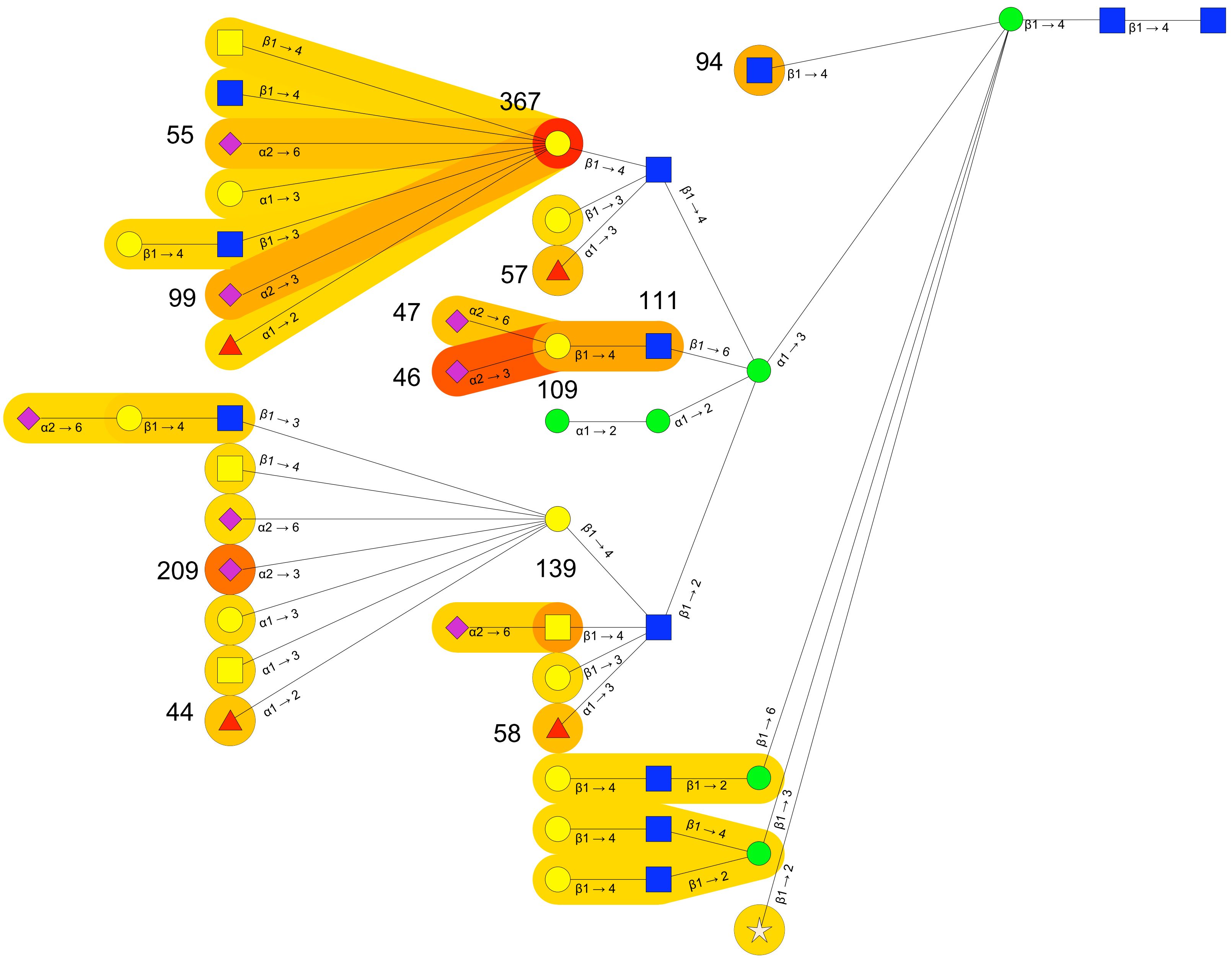


Figure 4.10: Continuation of composite structure comparison from Figure 4.9

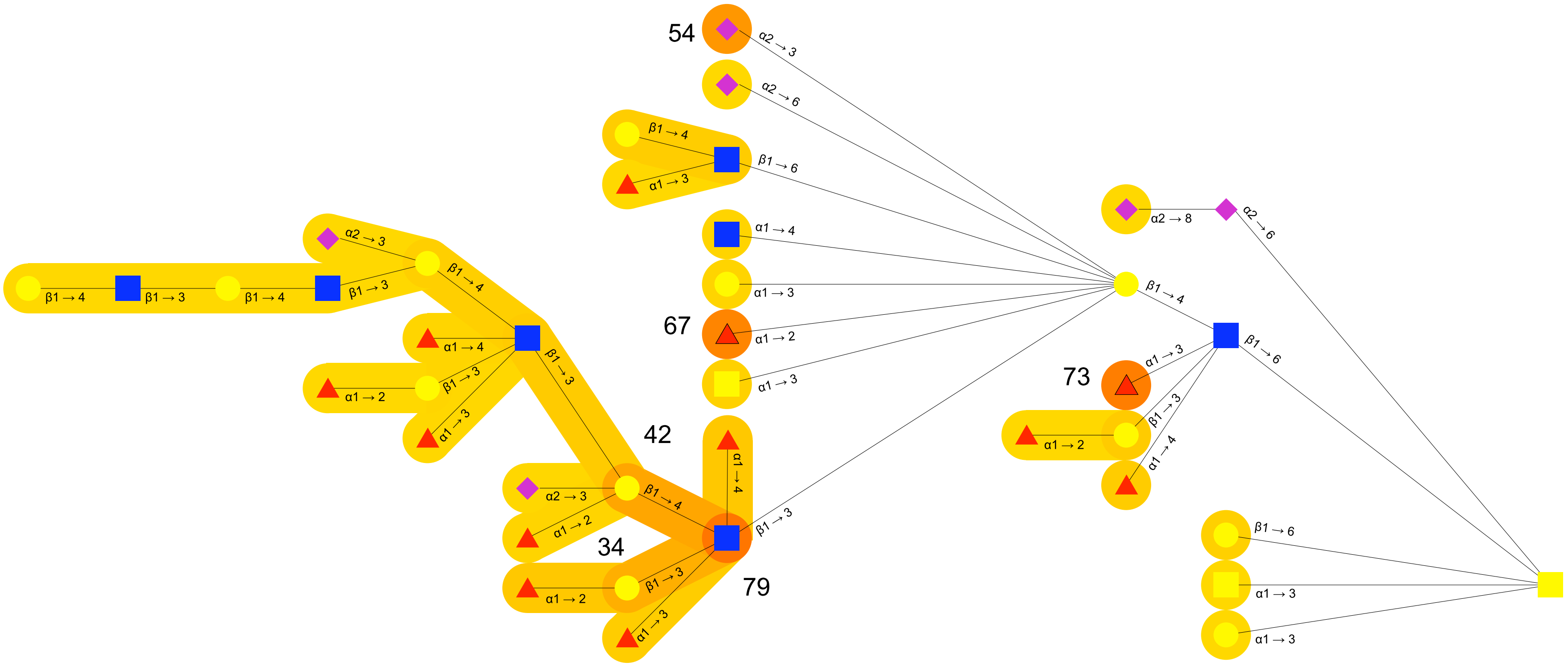


Figure 4.11: Comparison of composite structures from GlycoSuiteDB and GlycomeDB along the O-linked pathway (continued on Figure 4.12 and Figure 4.13). Differences in frequency of features are mapped by heat — red features being more commonly found in GlycomeDB. Residues without any heat information are those found in the pathways that are common to both databases. There are no structural features unique to GlycoSuiteDB.

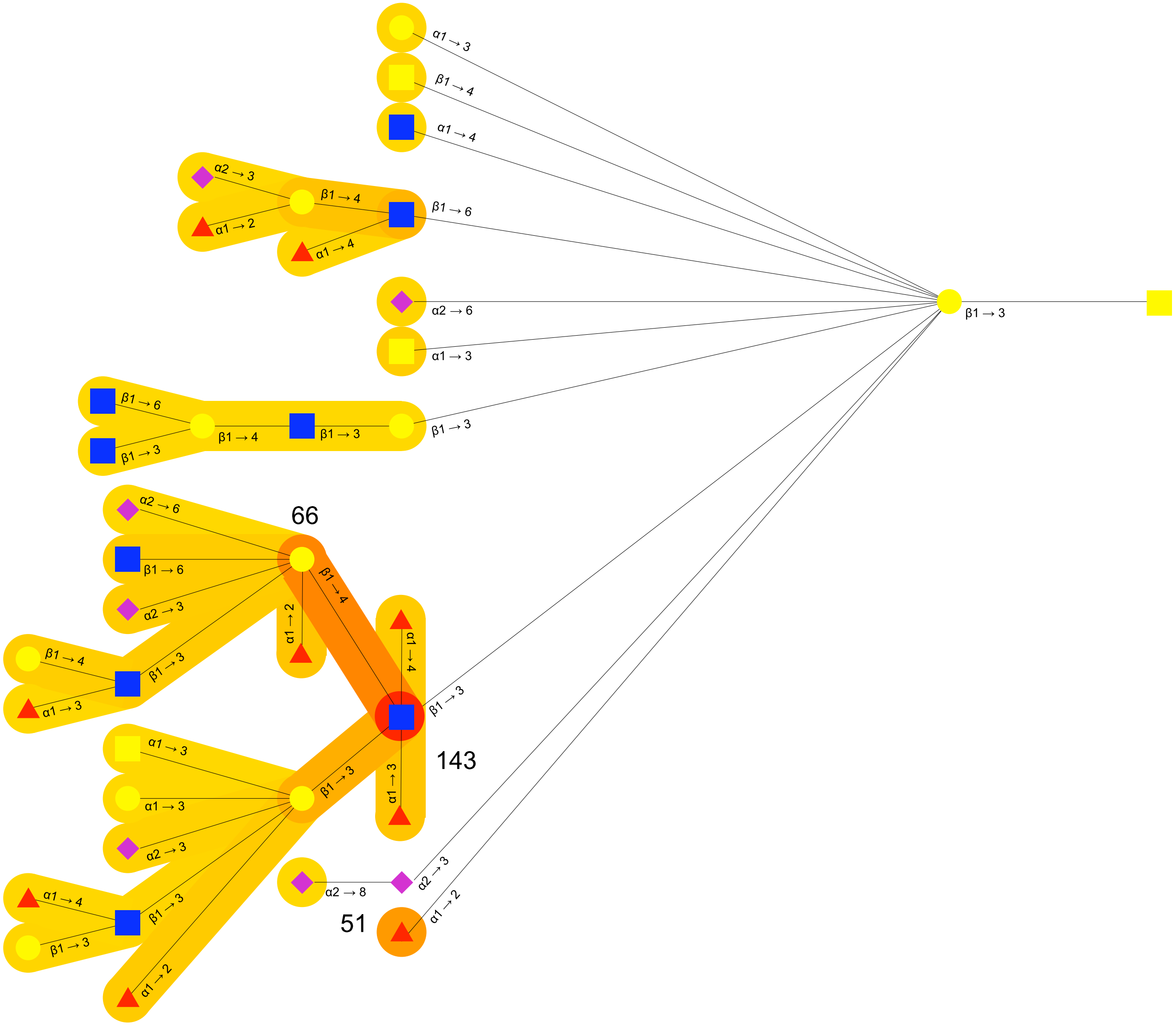


Figure 4.12: Continuation of composite structure comparison from Figure 4.11

In addition, the manual curation process used for GlycoSuiteDB allows for greater confidence in the correctness of the structures. Even so, structural features still exist that are clearly erroneous in nature. In order to maximise the number of structures which can be used, any structural features which are found only in GlycomeDB were only included if they were found in at least six unique structures. This threshold was chosen as it was the lowest possible number which still allowed for manual examination of outlying data points. For each of these composite structural maps, new structural features were resolved by studying the features, and determining whether they

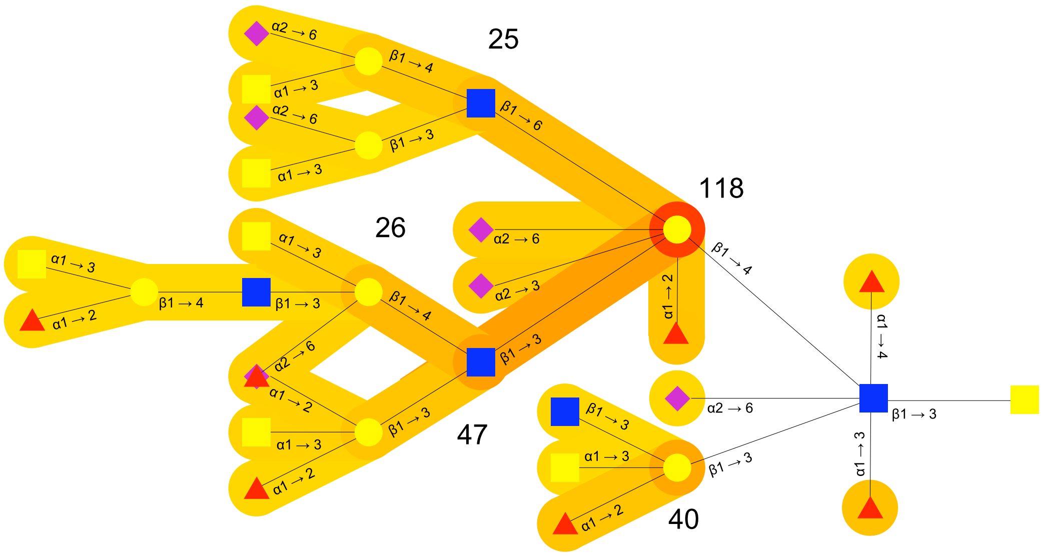


Figure 4.13: Continuation of composite structure comparison from Figure 4.11

represent new pathway data, or can be explained using current knowledge.

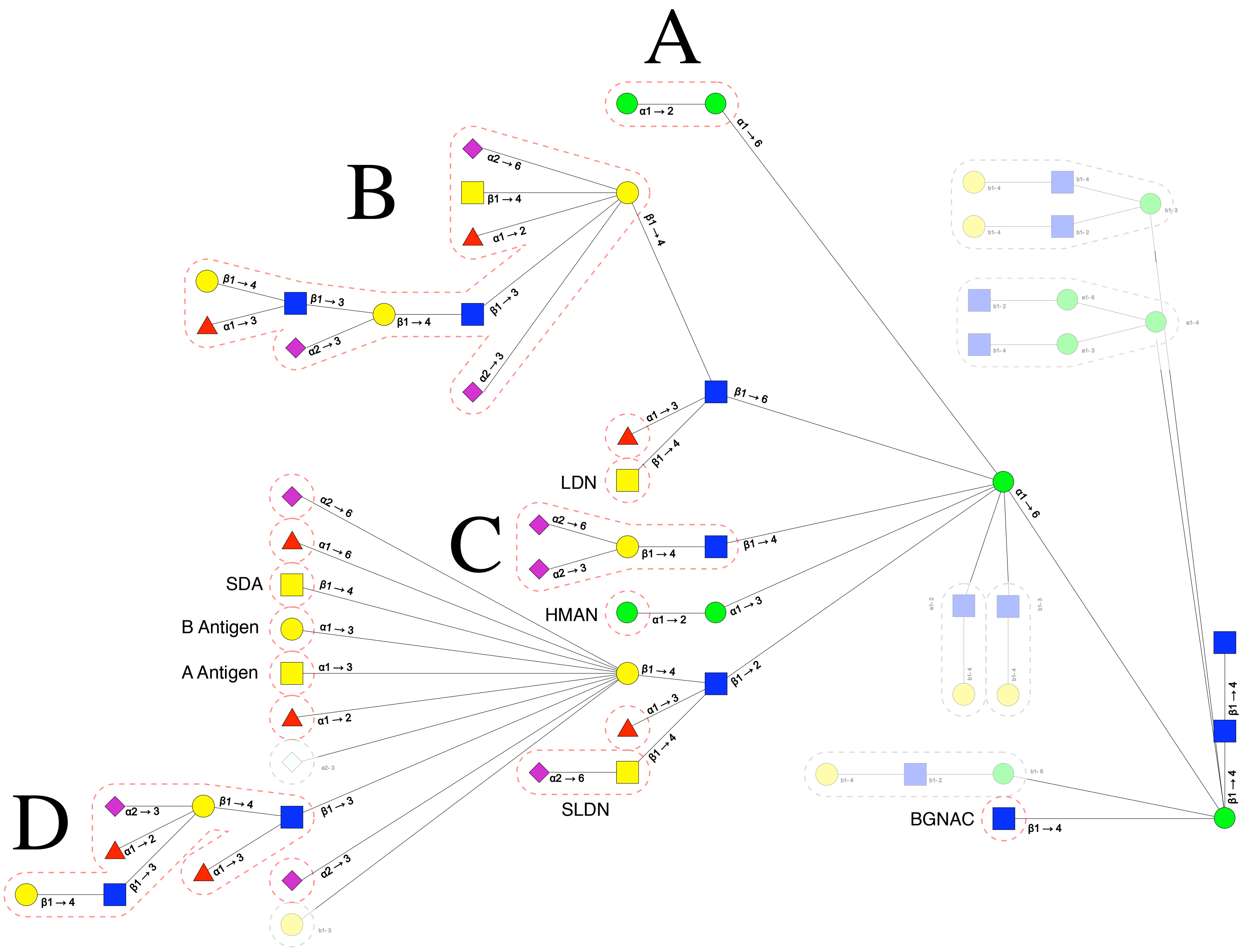


Figure 4.14: Additions along the N-linked pathway as marked out by a red outline. Larger structural additions are explained in the text, while individual residue extensions have been labelled appropriately as per the extensions key (Table 4.6).

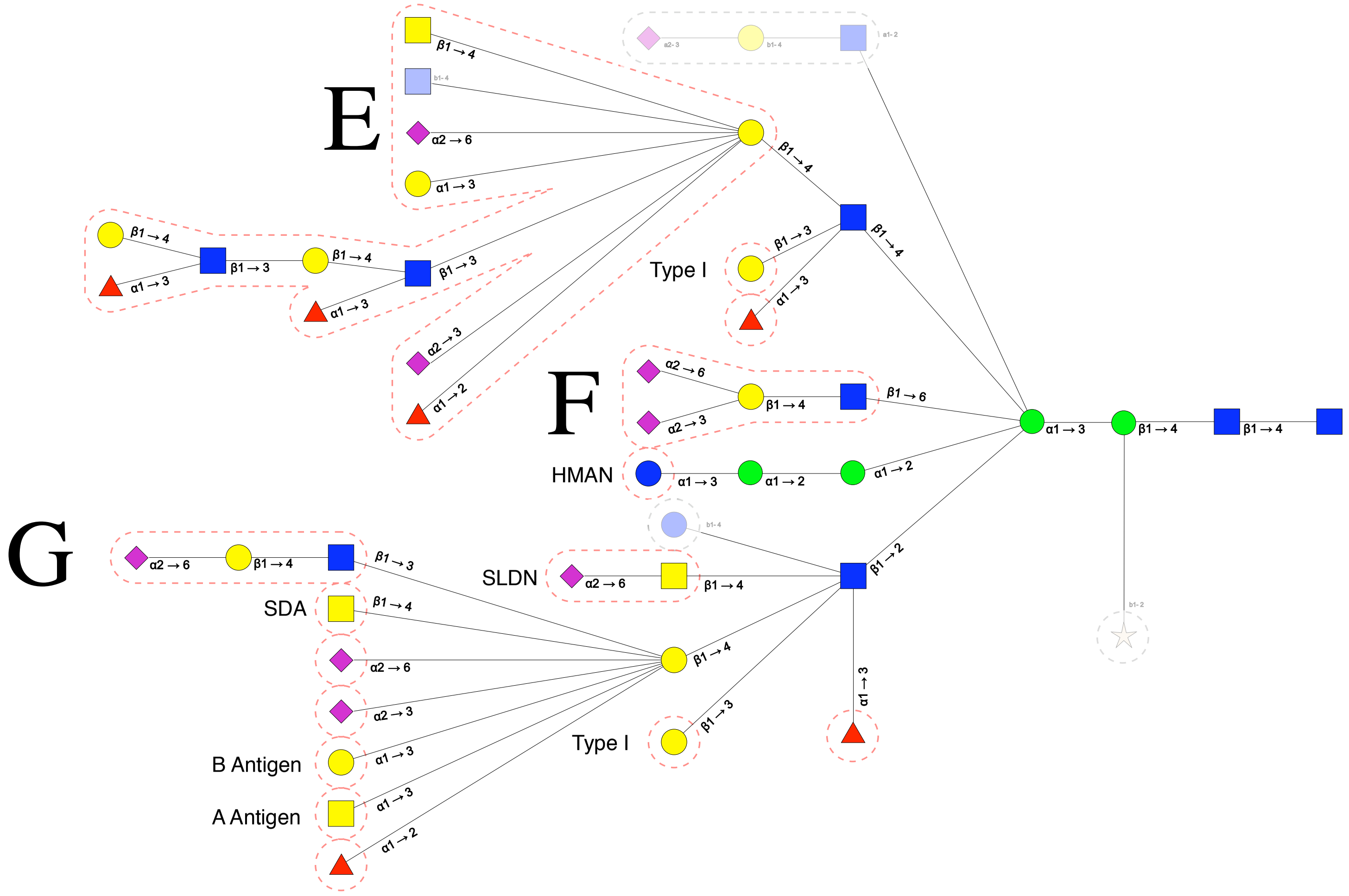


Figure 4.15: Continuation of additions along the N-linked pathway from Figure 4.14

4.3.1 N-linked pathway

The most complex of the composite structures, the composite N-linked pathway, contains 63 new structural features that cannot be resolved with the collected pathway information (Figure 4.14). A number of these features can be automatically discarded, since the linkages involved in the synthesis of the addition were found to be erroneous in the enzyme data collection phase. The additions to existing pathway information can be broadly classified into two groups: redundant pathway additions due to deficient pathway data collection, and pathway additions based upon structural motifs.

Additions due to deficient data collection are similar to those found for addition group A — describing a mannose disaccharide addition to the tri-mannosyl core N-linked structure. Since the KEGG data does not specify a specific ordering for the action of the glycosidases at the high mannose section in the pathway, it is possible that this particular structural feature was not taken into account. Similarly, the groups marked with HMAN can all be accounted for when manually referencing the KEGG pathways.

Structural motifs found in extensions are exemplified in addition groups B, C, D and E. Typical extensions include LacNAc chains, LacDiNAc chains, and the expression of the Sd$^{\alpha}$ epitope. The Sd$^{\alpha}$ epitopes are generally found only as additions to Gal residues at a level of six residues deep. LacDiNAc chains seem to occur only on extension of fifth-level GlcNAc.

Interestingly, a GlcNAc(β1→4)Gal disaccharide has been noted as an extension in group E. Further investigation into the structures that yield this feature show that referencing and database entry errors have yielded this feature (specifically structures referenced from [81] and [82]).

The Gal(β1→3)Gal extension cannot be found in any more than one reference for human structures, and so is not sufficiently evidenced to be categorised as a pathway addition.

The remaining Gal, GlcNAc and GalNAc extensions are Type I or Type II chains. The predominant chains seen are Type II chains, with the Sd$^{\alpha}$, A, B and O(H) epitopes being found at the shortest possible Type II chain.

α1→3 fucosylation in general occurs on any of the GlcNAc residues past the tri-mannosyl core. α1→2 fucosylation occurs on the Gal(β1→4) residues, and in a single case it uses as α1→6 linkage.

Sialylation seems to occur only as an extension to Gal residues. Six-sialylation occurring closer to the core is generally the rule for the structures deposited into the database. The exception to this general observation is the 6-linked structure seen in group G. There are only two papers which provide evidence for this structural feature [81, 83], both found in milk. The data here is not significant enough to draw any conclusions regarding the general location of these residues.

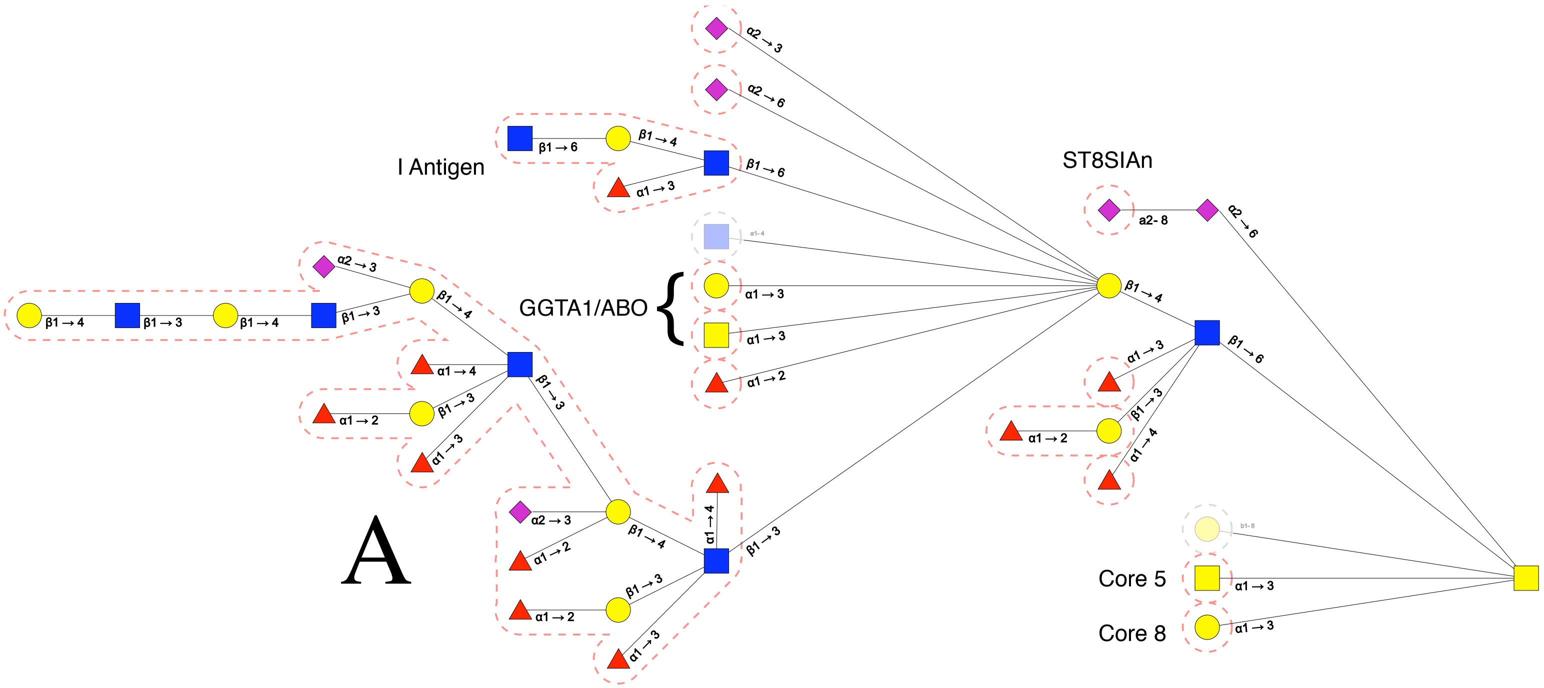


Figure 4.16: Additions along the O-linked pathway, as marked out by a red outline. For larger structural features, see the text for explanations, while smaller extensions are labelled as per the extensions key (Table 4.6).

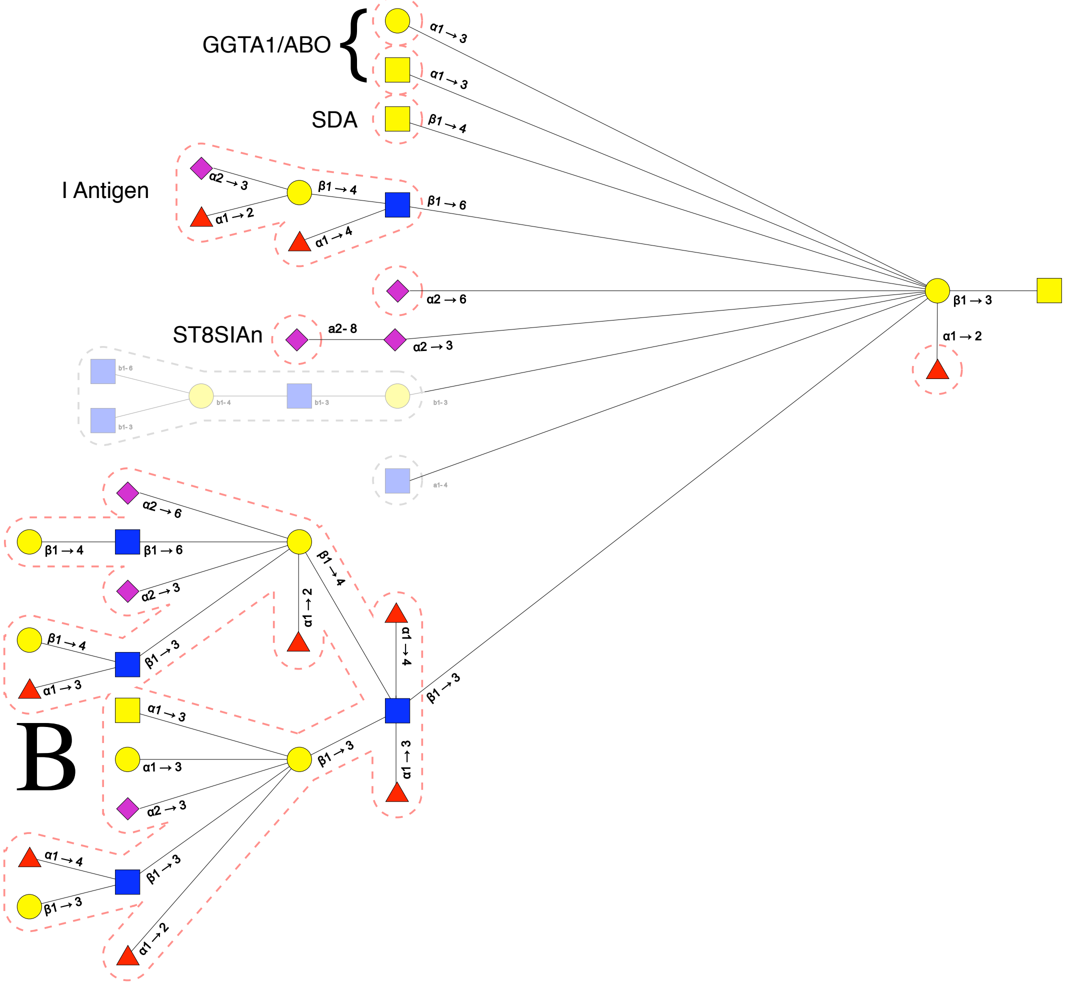


Figure 4.17: Continuation of additions along the O-linked pathway from Figure 4.16

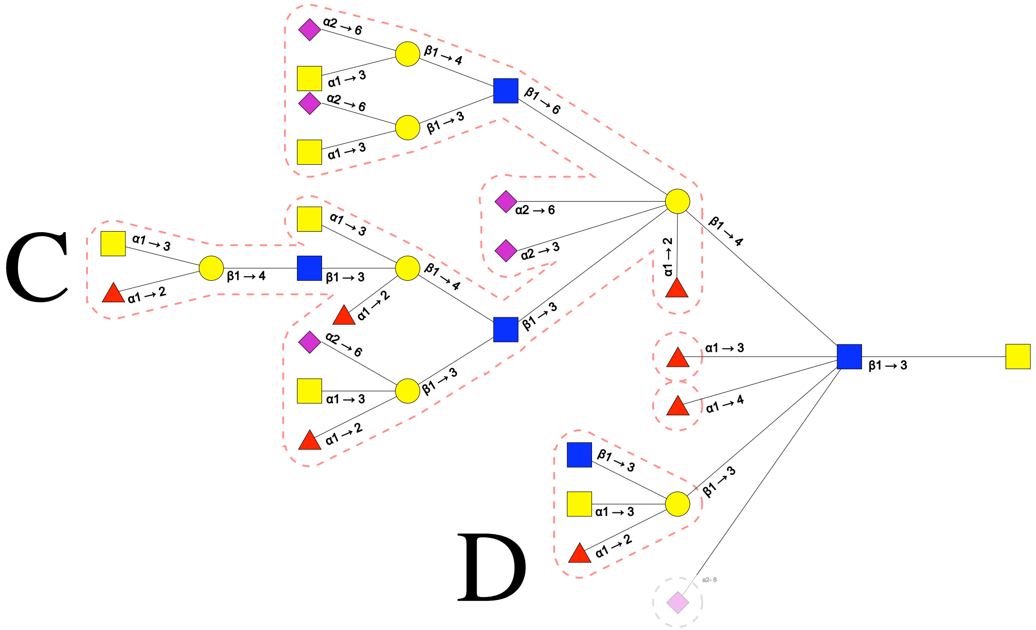


Figure 4.18: Continuation of additions along the O-linked pathway from Figure 4.16

4.3.2 O-linked pathway

O-linked glycosylation has been generally thought to be more complex and structurally diverse than N-linked glycosylation. Analysis of the pathway extensions on O-linked glycans (Figure 4.16) reveals the presence of certain patterns regarding the structure of the glycans. In general, extensions on the pathways are made up of Type I/II chain scaffolding, with the exhibition of Sd^{α}, A, B and O(H) epitopes.

Unlike the N-linked data, there are no additions which are already covered by the existing pathway. This is due to the more consistent nature of the pathway data obtained from KEGG.

GlcNAc(β1→6) linkages form I-antigen structures. The I-antigen extensions usually form branch points on the Gal residues. In the observed groups, the residues are only extended by the Gal(β1→4/3)GlcNAc(β1→6) disaccharide.

Extensions in group A are primarily Type I/II chain extensions. Although the Type I chains are less repeated than the Type II chains, there is no evidence to suggest that any Type I chains could not be repeated further. Group B consists of Type I/II chains, an I-antigen structure and Sd^{α} epitope. Group C can be constructed through a Type I/II chain framework, capped with A-antigen epitopes. Group D shows the beginning of a Type I chain capped with an A antigen. Any I-antigenic structures tend to yield shorter chains than regular Type I/II chains.

Fucosylation occurs upon the GlcNAc and Gal residues. Fucosylation on the GlcNAc residues occurs on 3- and 4-linkages upon Type I/II chains. Evidence for both types of fucosylation potentially occurring is limited in group B. Similarly, the pattern for fucosylation on GlcNAc is difficult to determine, due to the limited amount of evidence on these structural features. Fucosylation on Gal residues only occurs via an α1→2 linkage.

Sialylation occurs only on the Gal and GalNAc residues on 3- and 6-linkages. GalNAc sialylation only occurs on the reducing end GalNAc on a 6-linkage, as per the original pathway. Determining a pattern for the sialylation of Gal residues is not possible, as no obvious schemes for sialylation are evident. The absolute number of structures in which these features have been found is also low, so any conclusions drawn may be spurious. In addition, NeuAc residues are further substituted by NeuAc residues in two cases — both times on 6-linkages. The NeuAc(α2→6)GlcNAc residue has been determined to be a data entry error [84].

4.3.3 Glycosphingolipid biosynthesis

Additions along this pathway are fucosylations and Type I/II chain extensions. For the neo-lactoseries pathway (Figure 4.19), additions are predominantly for the A-antigen along the neo-lactose chain. Additionally, a fourth lactose unit is added to the chain. The Gal(β1→3)GlcNAc linkage forms part of a hybrid Type I/II chain.

Along the lactoseries pathway (Figure 4.20), the additions in group A are common to those found in the neo-lactoseries extension. Within group B, the extensions can be classified into Type

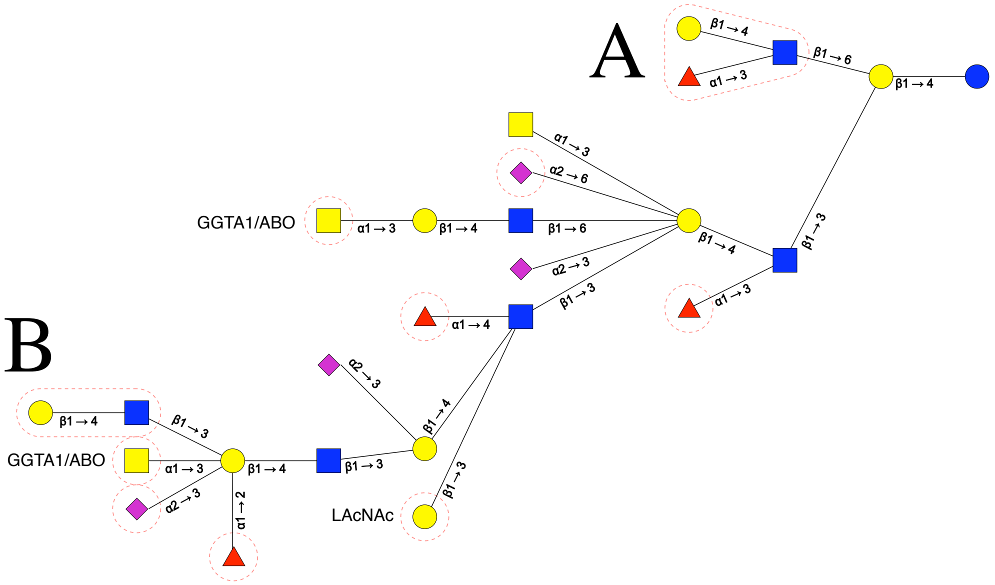


Figure 4.19: Neo-lactoseries glycosphingolipid extensions. Extensions are marked in red, and are explained in the text for larger extensions, while smaller extensions are labelled as per the extensions key (Table 4.6)

I and II chains. The Type II chains are regular repeating units, but the repeating Type I chain has been associated with carcinoma [85] [86] and has been shown on the Le(c) antigen.

4.3.4 Structures outside pathways

Structures that do not belong to any particular pathway also have features which need to be resolved. In order to identify the parts of the composite structure that are most important to study, a similar compositing of structures was undertaken, resulting in maps rooted at Gal, Fuc, GalNAc, Glc, GlcNAc and Man. Many of these structures exhibit characteristics of being sub-structures of larger structures, as they have the same structural features as their counterparts within pathways.

Gal-rooted structures (Figure 4.21) are all Type I/II chains as well as I-antigenic structures. Fucosylation and sialylation patterns are also similar on these structures, with fucosylation occurring only on the 3- and 4-linkage positions on the GlcNAc, the 2-linkage position on the galactose, and sialylation only on the 3-linkage position on the Gal residue. The Gal(β1→4)Gal linkage is not a well evidenced structural feature, and is an artefact due to an error in the database curation process [87].

Fuc-rooted structures (Figure 4.22) are not described in pathways, but the structures are known O-linked fucosylation structures. Possible enzymes for the synthesis of each of the linkages are marked out on the figure. It is important to note that the terminal NeuAc(α2→6)Gal linkage is the

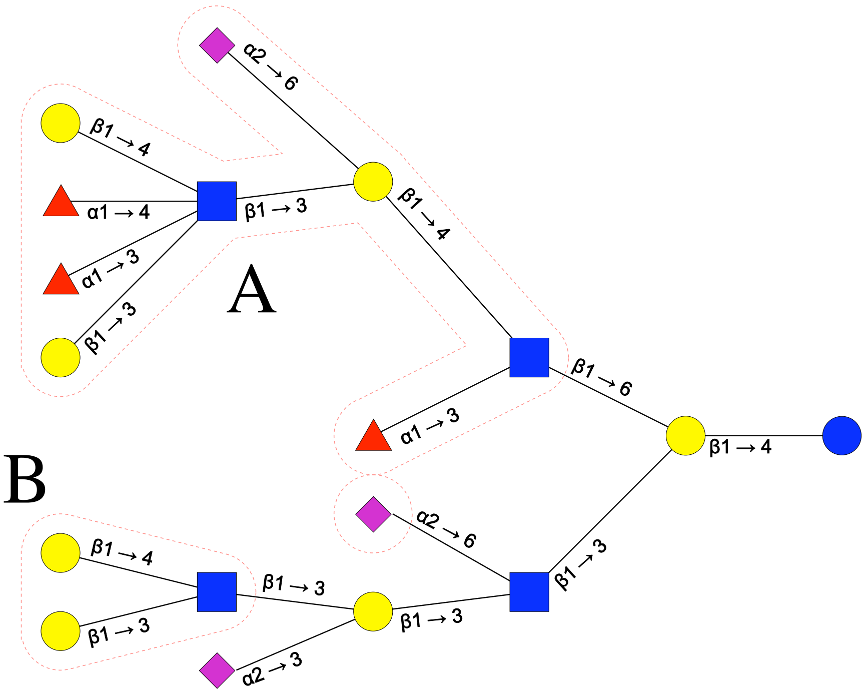


Figure 4.20: Lactoseries glycosphingolipid extensions. Extensions are marked in red, and are explained in the text for larger extensions, while smaller extensions are labelled as per the extensions key (Table 4.6).

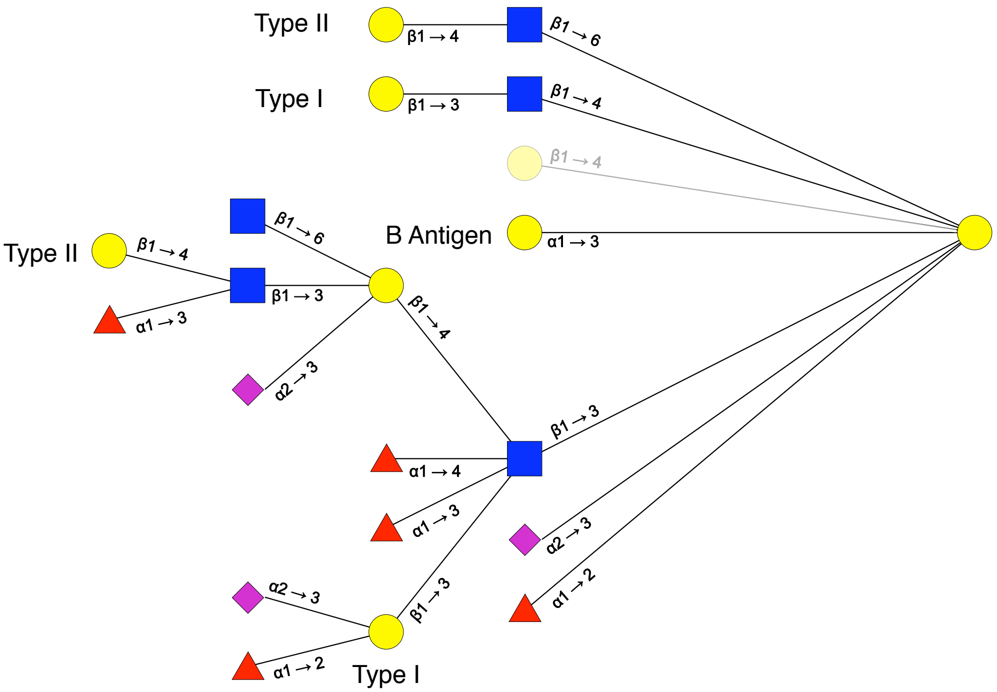


Figure 4.21: Additions to Gal-rooted sugars. Gal and GlcNAc extensions are labelled, while the Fuc and NeuAc extensions have not been individually labelled. Structural features that have been examined and found to be erroneous have been presented desaturated.

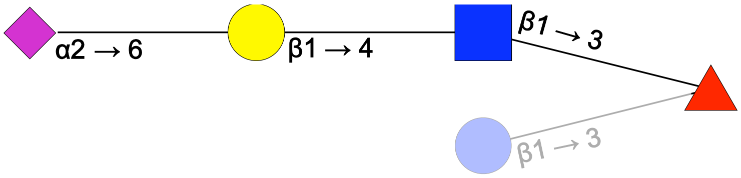


Figure 4.22: Additions to Fuc-rooted sugars. The Gal(β1→3)Fuc linkage has been examined and found not to be a valid new structural feature.

correct linkage position for this residue, as several reviews [88] [89] incorrectly list the linkage as a α2→3 linkage.

GalNAc-rooted structures are actually three sets of disaccharides: GalNAc(α1→3)GalNAc, GlcA(β1→3)GalNAc and Thr(α1→3)GalNAc. The first two can be resolved to fragments of the glycosphingolipid globoseries and chondroitin sulfate biosynthesis pathways with the genes GBGT1 and CHSY1 respectively. No enzyme exists in the database for the synthesis of the final disaccharide and it does not form part of any pathway described here.

Glc-rooted structures have a number of features. The Gal(β1→3)Glc linkage, as shown in the disaccharide analysis, may be an artefact. Structures such as fucosylated and di-fucosylated lactose are also observed. Most structural features are centred around modifications on a lactose core.

GlcNAc-rooted structures exhibit many similarities to the N-linked pathway, and most of the structural features can be traced back to the N-linked pathway. Exceptions are possible extensions upon LacNAc structures, with di-sialylation (NeuAc(α2→8)NeuAc(α2→3)Gal(β1→4)GlcNAc), and the B-antigen being found. The GlcA(β1→3)GlcNAc(β1→4)GlcA(β1→3)GlcNAc structure is hyaluronan, which is also found in a longer chain for the GlcA-rooted structure.

Man-rooted structures are all components of the N-linked pathway and do not require any further explanation.

4.3.5 Structures analysed

In total, 3348 human structures were analysed. This number does not include structurally ambiguous sequences, which were removed from the analysis due to their lack of credibility. Of the structures analysed, 57 repeating structures were not included in the automatic process, as it is more efficient to analyse them manually, rather than to introduce complexity into the program code.

Sulfates, phosphates and methylation were all removed from the remaining structures, as they can be considered secondary modifications to the foundation residues and sequences. In addition, 25 residue names were masked with a generic residue name, as the residue names could not be easily mapped onto the analysis model. The residue names were associated with various synthetic chemical modifications, and were deemed to have been synthesised apart from the regular biosynthetic pathway. As none of these residues were numerous enough to appear in the disaccharide table, it was safe to make this simplification.

Only 322 of the resulting structures fully matched the KEGG pathway data (Table 4.3). Given that there are 122 unique structures as end products in the pathways, it is clear that some structures have been duplicated due to stripping features such as sulfation and phosphates. None of the end products in the chondroitin sulfate biosynthesis pathway were found in the database. The 322 structures represent only 9.6% of all the structures in the database, which is a clear indicator of the degree of structural elaboration that occurs. In total, 844 structures did not match with a pathway, and the root residues of these structures are listed in Table 4.5.

Of the 2125 structures that partially matched with structures (pathway distribution shown in Table 4.4), the majority of structural additions could be found along the N- and O-linked pathways. This reflects the interest shown in the different classes of structure, and the attention devoted in literature.

Pathway	Number matched	Number in pathway
N-Linked	153	31
Neo-Lactoseries GSL	56	34
O-Linked	40	10
Ganglioseries GSL	28	20
Lactoseries GSL	27	10
Globoseries GSL	12	7
Chondroitin sulfate	4	6
Lacto and Neo-lacto series common GSL	2	3

Table 4.3: Count of fully matched structures per pathway.

Pathway	Number matched
N-Linked	1266
O-Linked	621
Neo-Lactoseries GSL	112
Lactoseries GSL	80
Ganglioseries GSL	22
Globoseries GSL	10
Lacto and Neo-lacto series common GSL	7
Chondroitin sulfate	7

Table 4.4: Count of partially matched structures per pathway.

4.3.6 Repeating structures

Repeating structures were examined separately. From a manual examination of the structures, repeating structures were identified as chondroitin, dermatan sulfate or chondroitin sulfate B, and as heparan sulfate structures. In addition, a poly-sialic structure was seen as found on NCAM [90]. A number of LacNAc(b1-6) repeating structures were seen, but after examining the literature, there were fewer than six unique structures which supported this structural feature, and so this

Root residue	Number of structures
Glc	109
Xyl	10
GalNAc	43
Gal	120
Fuc	4
GlcA	10
NeuAc	3
Man	47
GlcNAc	473
Other	25

Table 4.5: Root residues of unmatched structures.

Structural feature	Description
LDN	LacDiNAc structure
HMAN	High mannose structure
SLDN	Sialylated LacDiNAc structure
SDA	Sd$^{\alpha}$ epitope
B antigen	Part of Gal(α1→3)[Fuc(α1→2)]Gal epitope
A antigen	Part of GalNAc(α1→3)[Fuc(α1→2)]Gal epitope
BGNAC	Bisecting GlcNAc

Table 4.6: Extensions labelled in Figures 4.14, 4.16,4.19 and 4.20.

was deemed to be in error.

4.3.7 Summary of additions

Extensions on structures can be considered to exist upon the GlcNAc residues. There are a number of distinct extensions which can be applied to these residues.

Type I/II chain extensions In general, extensions upon GlcNAc are based upon an extension using a Type I/II chain, with branching occurring when a 6-linked GlcNAc is attached. Lists of the chains and the frequency of occurrence are listed in Tables 4.7, 4.8 and 4.9. I-antigen branching is only present in O-glycans, with the branching occurring mainly at the mannose core in N-glycans (with the exception of nine structures on N-glycans). Type I chains have been found as shorter repeating units (contradicting results from literature [91]), while the Type II chains can (theoretically) be indefinitely repeated. Chains of repeating Type II/6-linked GlcNAc features

Chain components	Count
II	3087
II,3,II	148
I	42
II,3,II,3,II	29
II,6,II	6
II,3,I	6
I,3,II	3
II,3	2
II,3,II,3,II,3,II	2
II,6	2
3,I	1
I,3,I	1
I,6,II	1

Table 4.7: Gal-GlcNAc chains found on N-linked structures. I/II corresponds to a Type I/II chain unit, while 3/6 corresponds to a 3-linked and 6-linked GlcNAc.

have been reported [92], although they have seldom been found in the databases.

Mixed Type I/II chains are present — but for mixed chains, the Type I units occur only at the start or the end of the chain. Similarly, 6-linked GlcNAc chain components can be mixed with the 3-linked GlcNAc, but only at the start or the end of a chain.

LacDiNAc structures A structural feature proximal to terminal extensions, the LacDiNAc structure (GalNAc(β1→4)GlcNAc) is not extended, but is found upon N-glycans on GlcNAc residues directly attached to the mannose core.

Fucosylation Gal residues exhibit only α1→2 fucosylation, while GlcNAc residues can be extended by 3- and 4-linked Fuc residues. Gal fucosylation occurs as part of a terminal epitope structure. The substrate analysis suggests that any GlcNAc fucosylation occurs after the Gal residue has been attached, and occurs in a linkage complementary to the chain; Type I chains receive α1→4 linked Fuc residues, while Type II chains receive α1→3 linked Fuc residues. This fucosylation pattern is common to all pathways.

Sialylation NeuAc residues can be used to terminate a Type I/II chain on an α2→3 or α2→6 linkage upon Gal residues. Since Gal residues are extended by GlcNAc residues on 3-substituted linkages, the 3-linked NeuAc is always a terminal extension. Substrate analysis of the 6-linked NeuAc indicates that the residue is found primarily as a terminal residue. The current data set

Chain components	Count
II	321
I	58
II,3,II	53
3,II	51
3	51
II,3,I	50
3,I	32
6,II	23
II,6,II	20
6	11
II,3,II,3,II	9
II,3	8
3,II,3,II	8
II,6	7
3,II,6,II	7
II,6,I	5
II,3,II,3,I	4
3,I,3,I	4
3,II,6	4
3,II,3	3
I,3	2
3,II,6,I	2
II,6,II,6,II,6	2
II,6,II,6	2
I,6,I	1
II,3,II,3,II,3,II	1
II,3,II,3,II,3,II,3,II	1
3,II,3,II,3,II,3,II	1
II,6,II,6,II,6,II,6	1
I,3,I	1
3,II,3,II,3,II,3,II,3,II	1
3,II,3,I	1

Table 4.8: Gal-GlcNAc chains found on O-linked structures. I/II corresponds to a Type I/II chain unit, while 3/6 corresponds to a 3-linked and 6-linked GlcNAc.

Chain components	Count
3,I	98
3,II	84
3,II,3,II	46
6,II	45
II	33
3,II,3,II,3,II	22
3,II,6,II	19
3,II,3,I	14
6,II,3,I	13
3	10
3,I,3,I	7
6,II,3,II	6
I	6
3,II,3,II,3,II,3,II	5
3,I,3,II	5
3,II,3,II,6,II	4
3,II,3	3
II,6,II	2
II,3,II	2
6,II,6,II	1
3,II,3,II,3,I	1
6	1
3,II,3,II,3	1
6,II,3	1
II,3	1
3,I,6,II	1
6,I	1

Table 4.9: Gal-GlcNAc chains found on Glc-rooted structures. I/II corresponds to a Type I/II chain unit, while 3/6 corresponds to a 3-linked and 6-linked GlcNAc.

seems to suggest that this linkage also is a terminal feature on chains, and not a mid-chain modification.

Epitopes In addition to regular NeuAc terminal residues, there are three chain terminating epitopes which can be found: Sd$^{\alpha}$, A-antigen and the B-antigen. The Sd$^{\alpha}$ epitope requires the presence of a NeuAc(α2→3)Gal substrate, while the A and B antigen are associated with the presence of the Fuc(α1→2)Gal substrate.

4.4 Structural validation

The pathway data has been verified to be a surrogate for evaluating sequences with respect to context. The implementation of the validation technique is a specialisation on the generalised algorithm for validating structures against pathway data (Figure 3.3). Any extensions on the structure to be validated will be compared against chain extensions, such that the chain must match the chain patterns listed above. Any structure that does not match these chain extensions will be deemed invalid.

I have implemented the algorithm in software as a web application. Figures 4.23, 4.24, 4.25 and 4.26 are screen-shots from the application. The validation tool is split into two components — the builder tool and the validator. The builder tool allows for the rapid and easy encoding of structures using a drag and drop methodology. The composed structure can then be sent to the validator tool that marks up the structure with annotation according to the rules elucidated in this thesis.

Both the builder and validator have been designed as modular components such that their functionality can be included in other applications. Using API encodings such as REpresentational State Transfer (REST), it is possible to call the application and retrieve the results as SVG or other image files.

Testing the performance of the validator, the software implementation was used to determine the number of structures fully covered by the algorithm from both the GlycomeDB and GlycoSuiteDB data sets. For GlycomeDB, 70 out of 472 unique structures (14.8%) were listed as containing errors. GlycoSuiteDB listed 27 out of 288 unique structures (9.3%). The structures chosen to validate both the databases were human structures stripped of any sulfation or phosphorylation, and any uncertain or fuzzy structures were discarded. This validation suggests that there are fewer errors in GlycoSuiteDB — a result consistent with the curation procedures that were put into place to maintain quality. It is not possible (without manually examining the data sets) to determine the false negative rate for the validation algorithm.

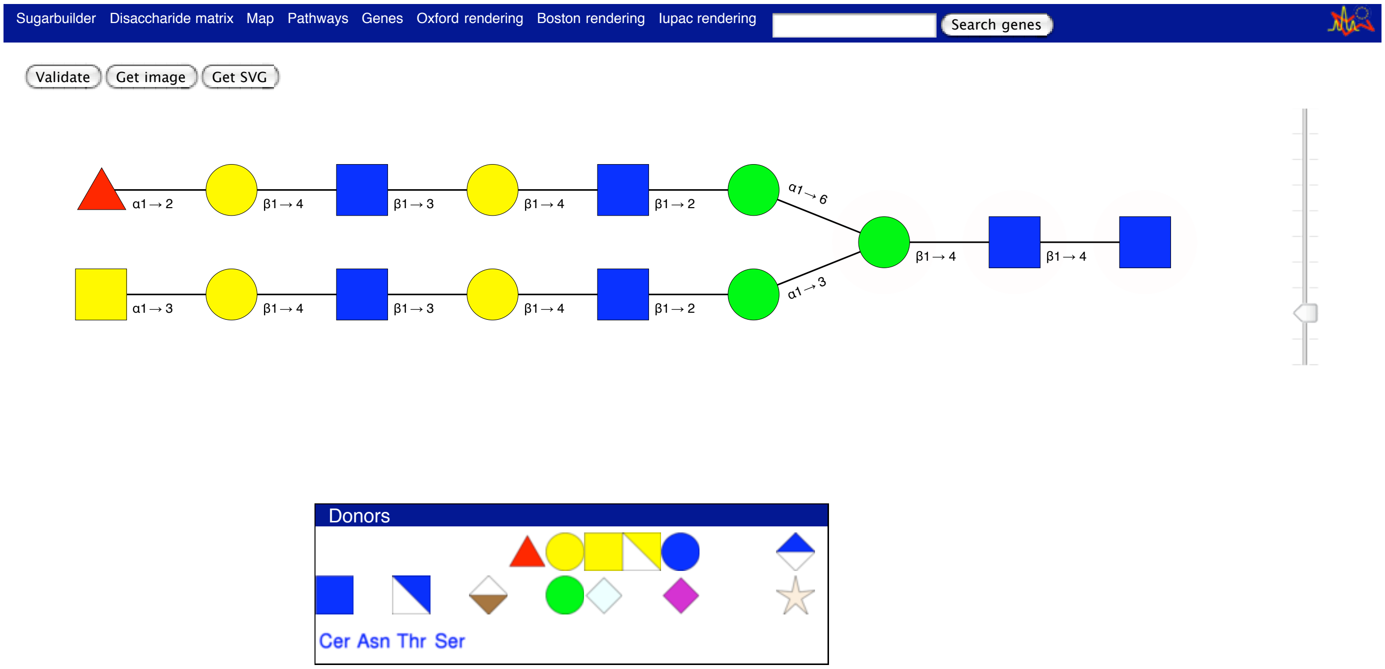


Figure 4.23: Screenshot of the sugar builder tool used to encode structures to be validated. Structures are incrementally built up using a drag and drop method.

4.5 Glycome estimation

When calculating the size of the glycome, we need to first estimate the size of chain extensions that are applied to the basic pathway structures. After estimating the size of chains, we can calculate the combinations of attachment points that the chains can be attached to, and have an estimate of the size of the human glycome.

4.5.1 Chain lengths

Enumeration of chain combinations for N-linked structures.

$$\begin{aligned} f(n) = \mathrm{II} &\quad + \quad n3\mathrm{II} \\ &\quad + \quad (n-1)3\mathrm{II} \\ f(2)\text{ also has these chains:} & \\ \mathrm{II} &\quad + \quad 6\mathrm{II} \\ \mathrm{I} &\quad + \quad 3\mathrm{I} \\ &\quad + \quad 6\mathrm{II} \end{aligned}$$

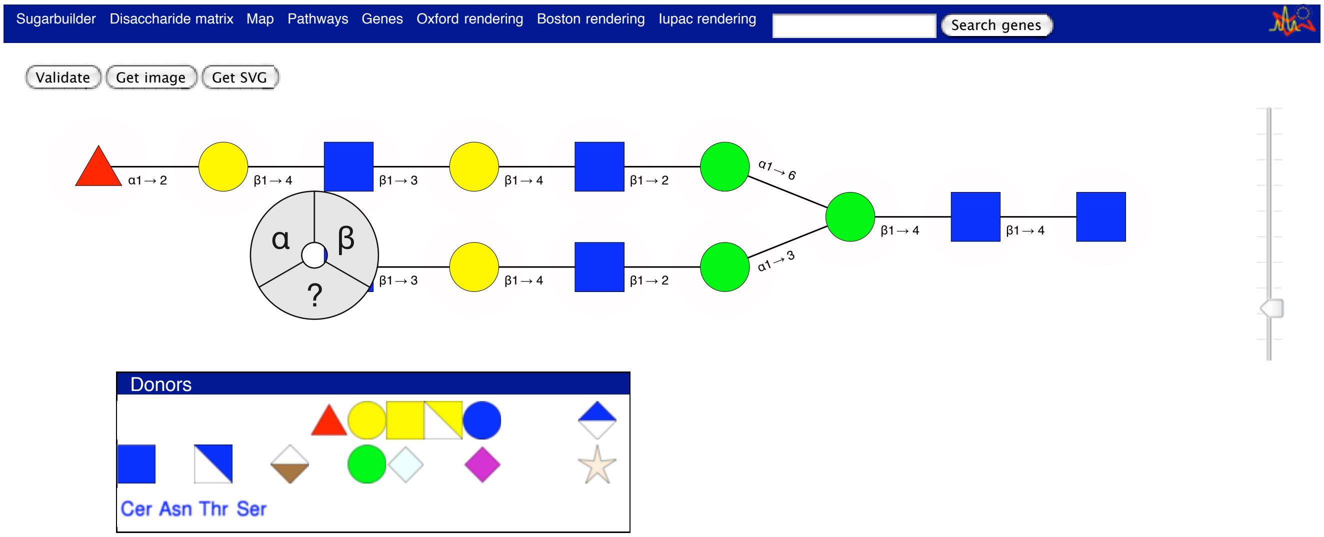


Figure 4.24: Anomer selection when building structures using the builder tool. A residue is selected from the donor palette and dropped onto the substrate residue. The anomer selector menu appears to allow the user to select the anomeric configuration.

$$+ \quad 3\mathrm{II}$$

Enumeration of chain combinations for O-linked structures is complicated by branching occurring along the chains. Since the branching for O-linked structures occurs primarily along the chain elaboration, the calculation of the number of chains is slightly modified. Branching occurs primarily at the start of the chain, and is not extended beyond a single unit.

$$\begin{aligned}
f(m,n) = \mathrm{I} \quad &+ \quad [m6\mathrm{I}] \ + (n-1)3\mathrm{I} \ (m = 0..1, n = 2) \\
&+ \quad [m6\mathrm{II}] \ + (n-1)3\mathrm{I} \ (m = 0..1, n = 2) \\
\mathrm{II} \quad &+ \quad [m6\mathrm{I}] \ + (n-1)3\mathrm{II} \ (m = 0..1) \\
&+ \quad [m6\mathrm{II}] \ + (n-1)3\mathrm{II} \ (m = 0..1) \\
&+ \quad [m6\mathrm{I}] \ + (n-2)3\mathrm{II} \ + 3\mathrm{I} \ (m = 0..1) \\
&+ \quad [m6\mathrm{II}] \ + (n-2)3\mathrm{II} \ + 3\mathrm{I} \ (m = 0..1) \\
&+ \quad (n-1)6\mathrm{II}
\end{aligned}$$

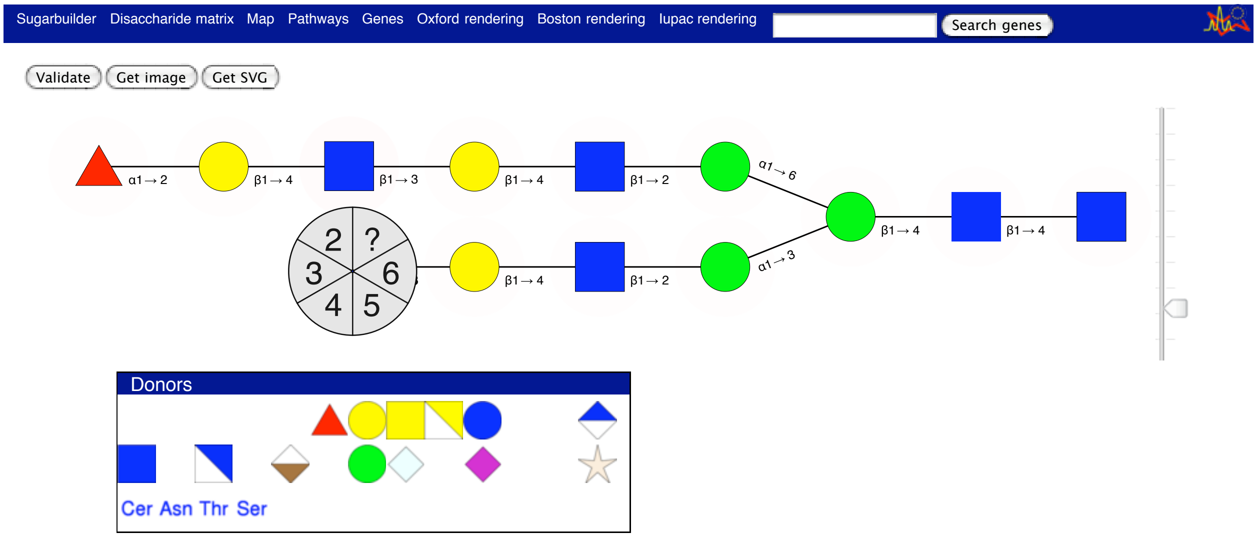


Figure 4.25: Linkage selection when building structures using the builder tool. After the anomer is selected, the linkage menu appears, allowing the user to select the desired substitution position on the substrate residue.

Enumeration of chain combinations for GSL structures.

$$n3\mathrm{II}$$

$$6\mathrm{I} + (n-1)3\mathrm{II}$$

$$3\mathrm{I} + (n-1)3\mathrm{II}$$

$$n6\mathrm{II}$$

$$6\mathrm{II}3\mathrm{II}$$

$$3\mathrm{I}$$

$$3\mathrm{I}3\mathrm{I}$$

Based upon the enumerations of chain types, a generalised formula for the number of combinations of chains can be derived. Since there are optional fucosylation events for each one of the chain types (out of 3I, 3II, 6I and 6II), there are 2^n different combinations of chain components — due largely to the number of ways in which the fucosylation can occur. O-linked chains are more complicated as the branching can occur along the chain itself. Generalising this to the different chain types:

N-linked chains $size(n) = 2^n + 2^{n-1} + 1 + 3 = 3(2^{n-1})$ $(+4$ when $n = 2)$

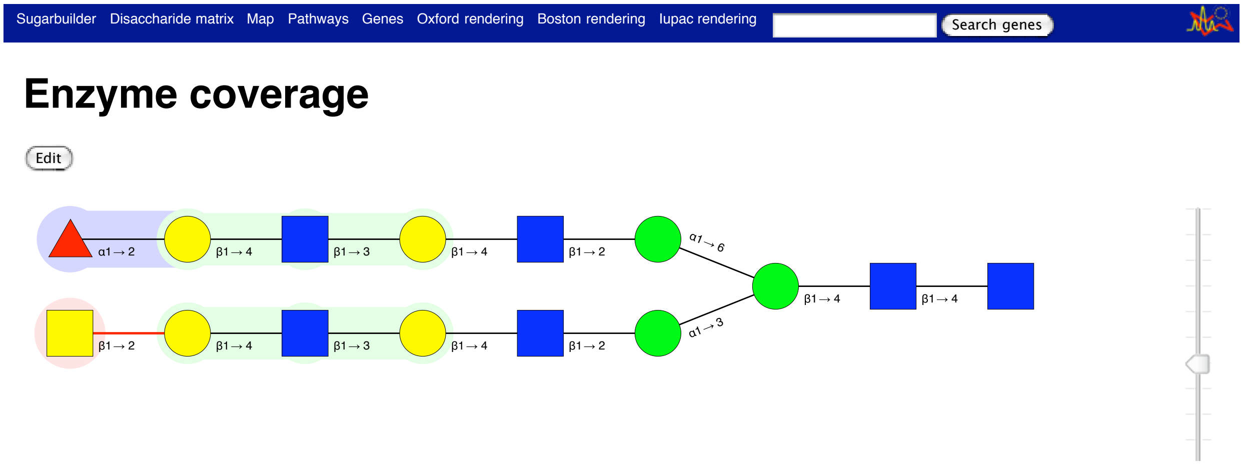


Figure 4.26: A structure that has had the validation algorithm run on it. This structure is not a valid structure - as indicated by the residues marked with a red highlight. Chain extensions are marked with a green background, valid terminal decorations are marked with a blue background, and invalid components are marked with red backgrounds. For invalid components, the linkages are drawn in red or blue, corresponding to no known enzymes and known enzymes encoding the linkage.

O-linked chains $size(n) = 4$ when $n = 1, 15 \times 2^n$ when $n = 2, 10 \times 2^n$ otherwise

GSL chains $size(n) = 2^n + 2^{n-1} + 2^{n-1} + 2^n + 3 = 6(2^{n-1})$ $(+2$ when $n = 2, +1$ when $n = 1)$

4.5.2 Extension sizes

Since the size of structures is limited, it is reasonable to examine the maximum dimensions of each of the extensions on each of the pathways. For each structure from the database, the extensions were collected, and statistics were collected on the number of extensions, as well as the number of residues and degree of branching within the extension (Table 4.10). The numbers listed cover 90% of the extensions observed. From the size of the extension, the maximum possible chain size that can fit into the number of residues is calculated.

4.5.3 Multiple chains

Since multiple chains can extend a structure, it is possible to calculate the total number of combinations of multiple chain lengths. For example, for two chain extensions on a single N-linked structure of length 1, there are a total of $(3(2^{1-1}))^2 = 3^2 = 9$ combinations. Two one-unit length

	% Total	Extensions	Size	Chain size	Branches
N linked	18	1	4	2	2
	34	2	3	1	1
	24	3	3	1	1
	24	4	3	1	1
	(of 2527)				
O linked	29	1	6	3	2
	46	2	4	2	1
	24	3	3	1	1
	1	4	3	1	1
	(of 1045)				
GSL	51	1	6	3	2
	40	2	4	2	2
	9	3	2	1	1
	(of 275)				

Table 4.10: Maximum sizes and branch counts to cover 90% of extensions for the given pathways. Extensions sizes are given in number of residues.

chains can be denoted as $1,1$. Extending this to the observed combinations of chain extensions, the total number of combinations for chain extensions can be calculated (Table 4.11). Not all combinations of chain length are listed, as not all combinations were observed in the data.

Multiple chain	Combinations
$1,1$	9
$1,2$	21
$2,2$	49
$1,1,1$	27
$1,1,2$	63
$2,2,2$	343
$1,1,1,1$	81

Table 4.11: Total numbers of theoretical chain extensions for N-linked structures.

4.5.4 Attachment points

To calculate the attachment points, the number of structures along the pathways that had extension points was counted. For example, 16 distinct extension points were counted for a single chain extension. For multiple chains, the ordering of the selection of attachment points is important, and so permutations (P^n_r) of attachment points are calculated. Examining the N-linked pathway elements, it is seen that for 2-chains, there are $3P^2_2+P^3_2+2P^4_2 = 36$ permutations of attachment points. For 3-chains there are $P^3_3 + 2P^4_3 = 15$ attachment points, and $2P^4_4 = 48$ points for 4-chains. There are seven 1-chain attachment points along the O-linked pathway, and only a total of four (two P^2_2 points) two-chain attachment attachment points. The higher number of extensions in comparison to extension points seen in Table 4.10 can be attributed to decorations such as fucosylation occurring on points along the existing structures.

4.5.5 Final calculation with elaborations

To calculate the final size, the terminal elaborations need to be applied to the chains. Examining the database, the A and B antigenic decorations are uniform — that is they are not found together in structures. As such, it is possible to consider there to be four types of modification A/B, Sd$^\alpha$, 3-linked NeuAc and 6-linked NeuAc. This assumption is further extended to all elaborations such that all chain elaborations are the same if present. In addition, there is a null modification, where there is no further elaboration to the chain. This means for each chain there are five possible extensions, which is multiplied by two for each further chain. Combining the attachment points and sizes of chains:

$$\begin{aligned}
total &= (1 - chain\ positions) \times (1 - chain\ combinations) \times 5 \\
&+ (2 - chain\ positions) \times (2 - chain\ combinations) \times 5 \times 2 \\
&+ (3 - chain\ positions) \times (3 - chain\ combinations) \times 5 \times 2 \times 2 \\
&+ (4 - chain\ positions) \times (4 - chain\ combinations) \times 5 \times 2 \times 2 \times 2
\end{aligned}$$

For N-linked structures, this is:

$$\begin{aligned}
total &= 16 \times 10 \times 5 + 36 \times 79 \times 10 + 15 \times 433 \times 20 + 8 \times 81 \times 40 \\
&= 800 + 28440 + 129900 + 25920 \\
&= 185060
\end{aligned}$$

For O-linked structures, this is:

$$\begin{aligned}
total &= 7 \times 72 \times 5 + 4 \times 3856 \times 10 \\
&= 2520 + 154240
\end{aligned}$$

$$= \quad 156760$$

In total, for N and O-linked structures, this yields a total potential glycome size of 341,820 distinct sequences. This is an approximation, as features such as degree of fucosylation and sialylation were not fully characterised and the branch length has been limited to cover the top 90% of most commonly occurring structural features. The estimate suggests that the total number of glycan sequence — at least in the human glycome — will be in the order of hundreds of thousands of structures.

4.6 Summary

Analysis of both individual enzyme data and pathway data suggests that the repertoire of glycan structures can be accounted for using a combination of pathway data and extensions. Generally, structures seem to be built up on a core framework, which is then decorated by various epitope, fucosylation and sialylation events. This observation confirms general knowledge about the synthesis of glycans. Although further modifications like sulfation and phosphorylation have not been accounted for here, the diversity of glycan structures appears to be restricted based upon observed structures. A description of the final algorithm for validating structures can be found in Section 4.4.

A total of 11 catalysed linkages were found in the database that did not have any enzyme information associated with them. Furthermore, three linkages catalysed by reactions associated with genes did not have any structures to support the existence of these linkages. See Table 4.1 for the list of linkages.

Substrate and neighbour analysis of the structures (Section 4.2) yielded no significant standalone insights into structure. While the neighbouring monosaccharide on a substrate did seem to correlate with known preferences and the order of application of transferases, further wet lab investigation is required to determine whether the suggested order of application of enzymes occurs in nature.

The validation algorithm was realised as a web application, and then used to determine the validity of the structures from both GlycomeDB and GlycoSuiteDB. While no conclusions can be drawn about the effectiveness or accuracy of the algorithm in detecting errors in the database, the analysis does provide some information as to the relative validity of GlycoSuiteDB over GlycomeDB.

An estimation of the size of the glycome yields approximately 350,000 distinct structures. While the calculation of the number of structures does not cover every single structural class, it does cover the classes most susceptible to exponential factors, and is a good approximation of an upper limit.

CHAPTER 5

DATABASE NETWORKING

ALGORITHMS FOR VALIDATING STRUCTURES were created as part of the EUROCarbDB project — a European effort to create a comprehensive database of experimentally supported glycan structures. The database uses an open-curation model to guide its data collection, and as such requires sophisticated automatic validation procedures to maintain quality.

One of the core goals of EUROCarbDB is to decentralise the database. This essentially means that the database acts as a distributed application, storing data locally to installations. The distribution of the database makes it less vulnerable to funding issues that are common with many databases. To enable this kind of infrastructure, a network layer that allows for the distribution of queries and raw data files was established during this thesis.

5.1 Application architecture

The EUROCarbDB application is designed as a web-based application, using a standard three-tier design with an added networking component. Standard architectural choices were made for the application to ease the development process, and to ensure that any changes to major components of the architecture would not have a significant impact on development.

The system can be generally categorised as comprising four key components — users, software, databases and the network. Figure 5.1 illustrates the relationships between the different components. The user interacts with the software either through an HTML interface, programmatically via the web services, or through a Java [93] application. Web services are offered using the SOAP protocol. Although the first prototype is a web-based application, the design has been chosen to accommodate the use of desktop applications as potential methods to deliver software. The software component interacts with the Tranche [94] network, experimental data and the backing database. Inter-node communication occurs via Tranche or through the web services interface. The system architecture is a variation on the three tier-design for web-based systems. The three-tier system encourages Model View Controller (MVC) design pattern usage, making the source code more maintainable and manageable. The MVC pattern can be mapped onto this design,

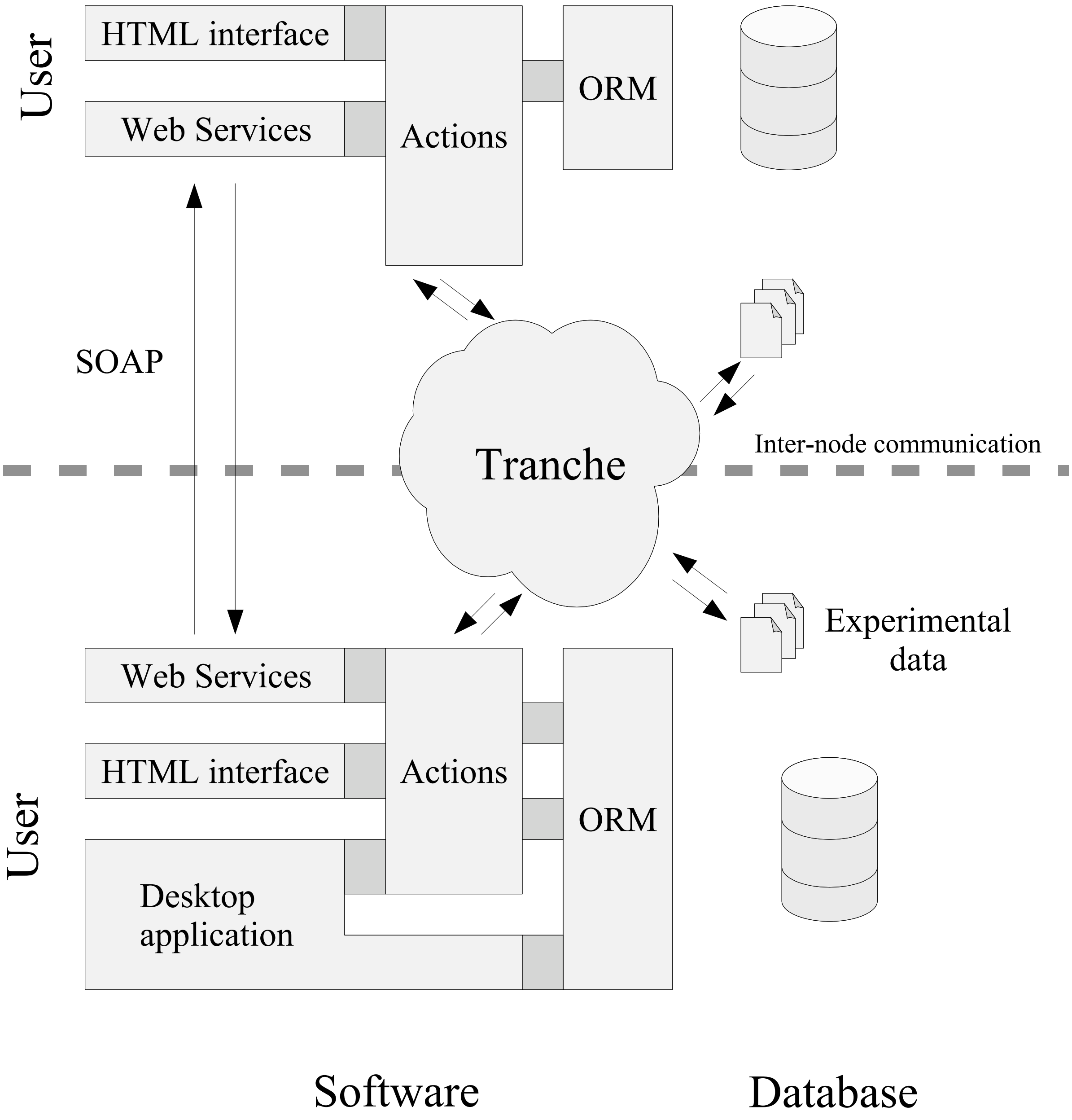


Figure 5.1: The overall architecture of two EUROCarbDB nodes communicating with each other. In the most simplistic representation, the system comprises software and database components. Interactions between nodes occur via SOAP or through the Tranche network cloud intermediary. Experimental data associated with records on each node can be transferred via the Tranche network.

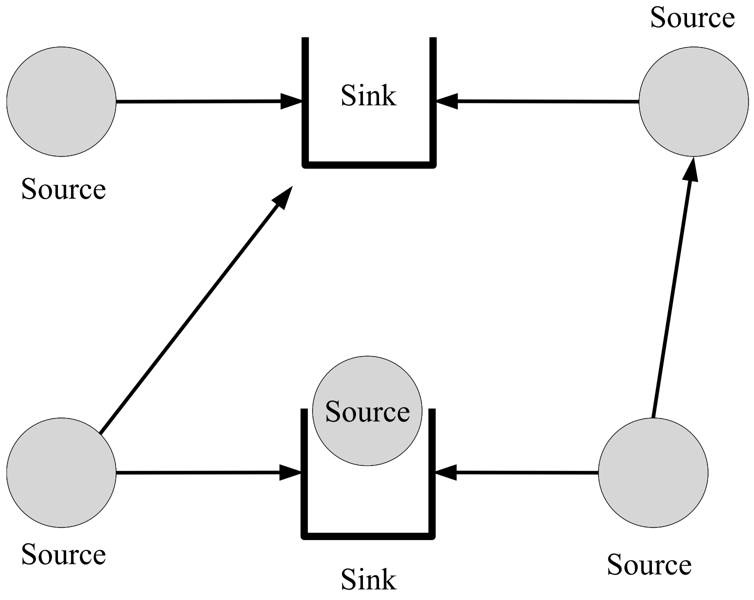


Figure 5.2: The overall network architecture of the system. Each source and sink is a node within the network, and the arrows represent interactions between nodes. Sources conceptually act as providers of information to the network, whereas sinks consume information on the network.

whereby the model is represented by the Object Relational Mapping (ORM) and database components, the view layer by the HTML and web services components, and the controller layer by the actions component.

5.2 Design

5.2.1 General network layout

To achieve the goal of distributing the operation of the database, a network layer is necessary. The goal of the network layer is to distribute the functions of the application transparently without user knowledge. The network is arranged as a set of nodes, with both pre-defined and *ad hoc* connections between nodes. Communication between nodes occurs on the query and data layers.

Nodes Each site participating in EUROCarbDB has at least one node on the network. Each node can have different levels of functionality, depending on the capabilities that the host institution wishes to expose. As shown in Figure 5.2, depending on the network layer in which the node participates, the node can be labelled as a data source, data sink or both. Source nodes provide data to the network via the data layer, whereas sink nodes provide query facilities over the query layer.

Node layout The network is laid out in an undirected mesh layout, with each node able to make connections directly with other nodes as required. The network is initially closed — with each node being aware of the other nodes *a priori* — but eventually expands to accept other nodes onto the network in a dynamic fashion.

The design of the network is split into two basic layers:

Query layer This layer allows queries and control signals to be distributed to the component nodes in the network and then handles the collation of results. The query layer has been designed for most ease when traversing firewalls and returns results in well-defined XML formats.

Data layer This layer allows the database to send raw data between nodes in the form of core database updates, or to distribute experimental data between nodes. The design of this layer is optimised for transferring large blocks of data at irregular intervals.

5.2.2 Query layer

The query layer is critical to allowing EUROCarbDB to function in a distributed manner. As well as allowing for distributed queries, this layer has several functions involving administration and signalling between nodes on the network. In general, the query layer provides a wrapper around the execution model of the application, exposing the actions as services. The query layer provides facilities to ensure the database can function without a central node, distributing queries to appropriate nodes in cases where the node itself cannot fulfil a query.

The use of a distributed query layer is needed when dealing with large data sets. Although facilities to transfer large data files from one node to the other are available, it is often faster to perform the query where the data is located. By distributing the query to the node closest to the data, the results can be returned more quickly as there is no additional data transfer required.

5.2.3 Data layer

The data layer is concerned with the distribution of primary data amongst nodes in the network — including the distribution of core data along the network and the distribution of experimental data amongst nodes.

The distribution of core data on the network is crucial to maintaining the integrity of the database across all the peers. Rather than requiring full availability of nodes to all other nodes on the network to maintain uniqueness through a query system, a synchronisation process has been designed to maintain copies of the core data across all nodes. Since an implementation of this synchronisation process has not been completed, details can be found in the 2007 annual report from the EUROCarbDB project [95].

5.3 Software results

The establishment of prototypes of the query and data layers is the key result of the EUROCarbDB developments. On the query layer, two prototypes were developed — a first implementation of database-to-database querying was established with the BCSDB [96] to investigate the difficulties that may arise when performing distributed queries. The main implementation resulted in the ability to distribute the action of arbitrary queries. This abstraction of queries is possible due to the design chosen. The data layer allowed for the upload and download of arbitrary files from the Tranche network. Although these two main achievements allow for some distribution of the application, further sophistication is required before it becomes fully decentralised.

5.3.1 Glycosciences.de to BCSDB connection

To test various protocols for connecting databases, the ability to send queries to other databases was tested on a connection between the Glycosciences.de database and the BCSDB [96]. The connection between the two databases was implemented as an add-on SOAP module within the existing PHP and Perl database applications. Although it was relatively simple to establish the API for remote procedure calls, complications arose from resolving the different data types that were stored in the database. Because different vocabularies are used, it became necessary to normalise them. Each query was forwarded from the spawning application synchronously — the source had to wait for the target to finish running its query before it could return results. This serialisation of queries has performance implications — but given the relatively low usage of the database, it is not practical to optimise the solution through asynchronous job dispatch.

5.3.2 Query layer implementation

The query layer implemented for EUROCarbDB acts as an adaptor on top of the action controllers. Through configuration of the application, it is possible to forward action requests to remote servers. This allows for remote dispatch of jobs. Results are returned from the remote actions, and then collated on the originating server. The interface is implemented using SOAP, serialising both query and results out to XML.

Actions are only required to be stateless to take advantage of this capability. Only input and output data fields are modelled in the SOAP messages, so any session or state information for a particular action is not saved. The execution of each action is like a single atomic unit of work, where all the relevant data can be encapsulated in the input message and output result. The system was tested using some simple actions, and was further used to facilitate a connection between a remote application (Glycoworkbench) and the server application.

No stress testing of the application was undertaken. Since there is neither a significant amount of data in the database, nor a sufficient number of nodes to test scalability, it is difficult to test the performance of the system under heavy load. In regular usage, the system itself will only see light usage, so in many respects the performance of the system under stress conditions is not important.

5.3.3 Data layer implementation

The data synchronisation layer has not been fully implemented in EUROCarbDB. To achieve some degree of data synchronisation, the Tranche system has been integrated into the application.

To integrate Tranche into EUROCarbDB, the separate upload, download and daemon software modules were integrated into the web application framework. Configuration data common to the modules was integrated into the core application configuration, and the daemon process tied into server startup and shutdown. Users are converted into Tranche users from application users automatically, with the user identities signed with the appropriate certificate.

The basics of data synchronisation have been added to the application – data can be serialised and deserialised, but no record merging is in place to handle updates of records.

5.4 Implementation details

The application is implemented as a first prototype using a three-tier application environment — operating on the Java [93] platform, using the Hibernate [97], Struts 2 [98], Tomcat [99] and PostgreSQL [100] libraries to add the desired functionality to the basic server. In addition, JiBX [101] as well as XFire [102] are used to supply the XML functionality. In general, open-source libraries and applications are used as the basis for the entire application. An overview of the general software application architecture can be seen in Figure 5.3.

Java and Tomcat

The Java language is a bytecode-compiled language, known for the ease of cross-platform development and wide availability of libraries. The Apache Tomcat application server provides an environment under which Java web-based applications can be offered to the public. Both the Tomcat server and Java are open source, albeit licensed under different schemes (Apache 2.0 License and GNU Public License (GPL) v2 respectively).

Hibernate

Hibernate provides an ORM layer onto the database by reverse engineering a set of Java source-code files from the database schema. By reverse engineering the Java API from the database

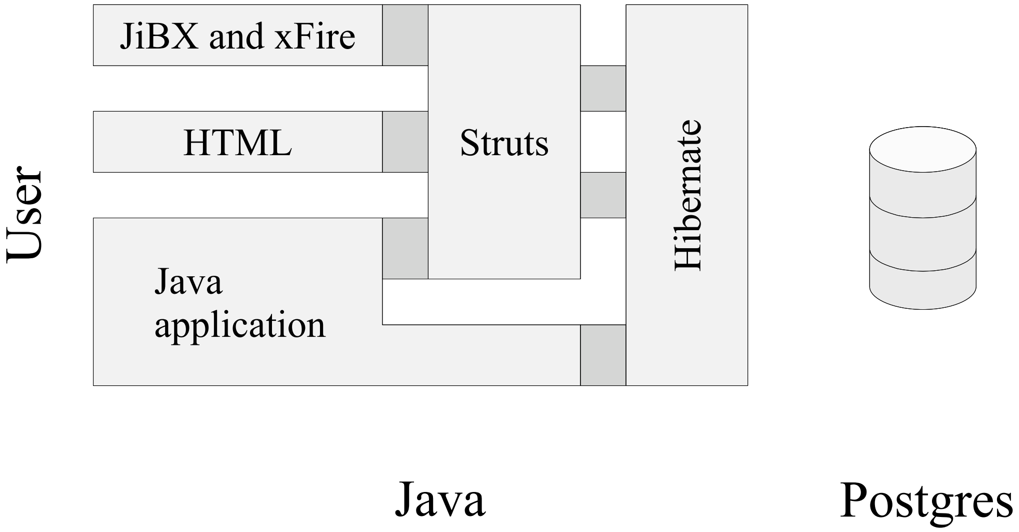


Figure 5.3: Relationships between the software components in EUROCarbDB. This diagram illustrates the software modules for the various architectural features shown in Figure 5.1. The architecture is a Java [93] and SQL database-based system, in which the Struts and Hibernate components form significant reused parts in the architecture.

schema, there is no need to synchronise the changes between system models, the database model and object model. As the reverse engineering process is a somewhat naive process, the reverse engineered source code needs to be modified so that an appropriate API is exposed for manipulating the objects.

PostgreSQL

PostgreSQL is used as the database component in the application. The optimal way to store data for the application is in a relational database, and the PostgreSQL database provides the best set of features and performance of the open source databases. Connection to the ORM component, and the Java applications in general is achieved through the use of JDBC libraries.

Struts 2

Web requests are essentially stateless operations. Each request does not know about the requests which have occurred before it, and there is nothing within HyperText Transfer Protocol (HTTP) to explicitly bind a series of requests together in a transaction. For this reason, support for saving state has to be added to the server or client side. Server-side state is established by using cookies — identifiers stored on client machines — which can tie a particular request to a particular server state. Many changes to state can occur for a request, such as a user logging in, or being part-way through a multi-stage action. Because much of this functionality is common to many web applications, a number of frameworks have been developed to provide these common functions. One such framework is Struts. Struts follows an MVC approach to the development of web applications, providing a set of Java classes which provide the basic functionality within the application.

Core to the function of Struts is the concept of an action. Source code associated with an action is related to a single unit of work, or a verb, which is applied to the system. Simple actions can be 'Add biological context', 'Associate taxonomy', or 'Upload experimental data'. Each action should be atomic, and should not be dependent on the state of other actions. Struts manages the execution of the actions, as well as executing a series of pre- and post-execution methods as interceptors.

XML libraries

JiBX and XFire together provide support for SOAP within EUROCarbDB. In order to avoid duplicating work, JiBX was used as the binding library for both SOAP and generic XML marshalling and unmarshalling. JiBX functions as a bytecode manipulation of classes, marking out the binding of fields in an object to elements in an XML file. The XFire library is an implementation of the SOAP protocol, and handles the dispatch of SOAP messages to appropriate methods. A specialised SOAP message to Struts' action wrapper was written to allow the use of all actions as SOAP methods.

5.5 Conclusions and future directions

The development of a comprehensive database for glycomics is of great importance to the continued development of the field. While the collection of data can help in the short term, the long-term benefits through data mining are even more attractive. The EUROCarbDB project, of which this thesis is a part, aims to meet the needs of the glycomics community. To meet the goals of distribution, accessibility, longevity and quality, a network component was developed as part of the project to lend decentralised qualities to the software. The network component was developed in two parts: the establishment of a query layer, and the integration of the Tranche system to distribute large files.

The creation of a query layer was undertaken to allow the distribution of queries between nodes on the network. As it may not be expedient to transfer the raw data from node to node when *ad hoc* queries are performed, the application needs ways to send queries to other nodes and collate results. A SOAP-based API was used to interface the controller logic of the application, acting as a proxy to remote execution of the controller. The system was tested using stateless actions, as the nature of the proxying did not lend itself to stateful remote execution.

Although the data layer was fully designed, its full implementation was not completely established for the thesis. The Tranche system was integrated into the project to allow for large data sets to be easily transported from node to node. The integration featured both user and daemon integration.

Future development of the network layer will involve the implementation of database synchronisation. Because the network has been tested neither under load nor in a scaled-up network topology, the performance of the system under these stress conditions will need to be evaluated and optimised.

Chapter 6

≈

Conclusions

Maintenance of distributed databases is a challenge due to the decentralised nature of the data. To ease the burden, tools can be designed to automatically check the data for both syntax and context validity. Whereas syntax-checking algorithms can be independently applied to incoming data, context-checking algorithms are necessarily dependent on examination of other entries within the same context.

Validation of structures and sequences should be done at the time of data deposition to provide the most timely feedback to the depositor. To implement this kind of checking on a distributed database requires the examination of the full data set. In practical terms, this means that the deposited sequence should be compared against the complete glycome to examine whether the sequence contains any new structural features that need to be further verified. Because no complete databases of glycomes exist, a complete contextual examination via one-to-one sequence comparison is impossible. Instead, biosynthetic information can be used to provide contextual data against which structures are verified.

Since all glycan structures follow a biosynthetic pathway that guides synthesis, all sequences within a context must conform to the rules of synthesis. The degree to which a structure conforms to the biosynthetic rules can be determined by a simple test which acts as a surrogate to test the appropriateness of the sequence to the full context. Any structural features that are not controlled by a biosynthetic pathway are either errors or represent new biosynthetic knowledge.

Before verifying new sequences, the pathway data was validated to ensure the quality of the reference data. To achieve this, pathway data was compared with existing sequence databases to clearly identify potential paths for elaboration of structure.

In the process of validating enzymatic data, several ancillary results were found. By examining sibling residues, enzymatic complexes or any possible substrate specificity for an enzyme were identified. Although this analysis did not provide definitive identification of these properties, it served to identify potentially interesting reactions to be examined further through wet lab procedures.

A comprehensive list of glycosyltransferases was obtained through the verification process, and several new reactions without enzymatic data were identified. Linkages that resulted from errors in data entry were also identified.

The verification process yielded a matrix of verified glycosyltransferases (Section A.1). This matrix contains several entries that are new to existing reaction matrices, and are indicative of deficiencies in the summaries of data already provided [11]. The 11 reactions not associated with enzymatic information must be investigated through further experimental methods.

As enzymes were verified, linkages and their associated reactions were deemed to be artefacts if the reference evidence did not support the presence of these linkages. The volume of eliminated linkages highlights the large number of errors found in existing databases.

To realise the verification and validation of data, software was developed in the Ruby language as a web application. In addition, a software toolkit was developed to allow for varied analyses to be performed on sequence data originating from many sources.

Pathway validation — the process of checking biosynthetic pathway data against the database — resulted in the confirmation that the pathways from KEGG are representative of the structures in the database. In addition, extensions to the pathways were identified, comprising Type I/II chain extensions, terminal epitope elaborations and fucosylation. Patterns in chain formation were observed, which limits the number of structures that can be formed. Information about chain length, degree of fucosylation and degree of branching was not established in this analysis. From the information gathered, it is possible to determine whether a structure has a core structural component that exists in known pathways, and which extensions have been combined to result in the end structure.

The results shown here indicate—for human structures, at least — that there is an upper limit on the diversity of glycans, and that their synthesis is regulated strongly by known biosynthetic pathways. This limitation on diversity permits algorithms to be developed that can be used to verify structures deposited into databases. This will facilitate both syntax and context based validation of sequences, significantly reducing the difficulty involved in maintaining the database.

6.1 Structural diversity

A significant conclusion from this study is that there is an upper limit to the diversity of human glycans, significantly lower than some previous liberal estimates of the size of the human glycome. Structures can be grouped into sets of elaborations on core structures with Gal → GlcNAc linkage repeating extension units, and various structural epitopes or residues at terminal positions. A number of new disaccharides have been identified, suggesting the presence of further glycosyltransferases which could synthesise these linkages. No position along the pathway for these enzymes has been proposed. Calculations to estimate the size of the human glycome, based on the listed model of glycosylation, suggest an estimated 340,000 glycan structures. Numerous flaws in the model were identified, and further data mining is needed to refine the model for a more accurate estimate of the size of the glycome.

6.2 Structural validation tool

Using a codification of the structural features and pathway information elucidated in this study, it is possible to write a software tool which will permit the verification of structures to gauge their contextual relevance to human structures. As an exemplar of such an endeavour, the validation algorithm was implemented as a web-application for the validation of structures entered using a specialised input tool. This has limited general applicability, since glycan databases typically contain structures from a wide variety of species. For a verification tool to be generally valid, a similar analysis must be performed for each system to ensure that the pathway information is valid.

6.2.1 EUROCarbDB network layer

The validation methodology was developed for the EUROCarbDB project. The validation allows for open curation of the database to be more easily managed, and allows its quality to be maintained despite the absence of centralised management. The decentralisation of the database has also been facilitated in this thesis by the establishment of a network layer within the project. The two major results from this work are that queries on the database can be transparently distributed to other databases on the network, and that facilities for the transfer of large data sets are now built into the EUROCarbDB application.

6.3 Future directions

This analysis rests on a fundamental and logical leap of faith: That the sequences studied and deposited into databases are in some way representative of the entire spectrum of the human glycome. Verifying this is impossible without performing a concerted sequencing effort for the entire human glycome. Any further improvement on this work will not come through pure bioinformatics — but rather will be realised through an effort to provide complete sets of glycomes for further study. This requires more efficient high-throughput analytical methods, which in turn requires better bioinformatics to analyse the resulting data.

There exist tangential applications for extension of the work surrounding the study of glycan diversity. Areas such as the automatic sequencing and analysis of glycan profiling and mass spectrometric data could benefit from such insights. An automatic algorithm for detection of glycan structures from mass spectrometric data is feasible, combining structural information from pathway data with spectral data to limit the combinatorial explosion inherent in de-novo sequencing. It should also be possible to design tests for spectral data to find peaks that do not agree with pathway data, and are thus indicative of fundamental differences in the glycosylation machinery — allowing one to skip over the core logical fallacy inherent in this analysis, and discover new structural features.

Appendix A

Additional data

A.1 Human glycosyltransferases collected

Reaction	**Genes**
Gal(β1→4) ⇒ GlcNAc	B4GALT1, B4GALT2, B4GALT3, B4GALT4, B4GALT5, B4GALT6
Gal(α1→4) ⇒ Gal	A4GALT
Gal(β1→4) ⇒ Glc	B4GALT6
Gal(β1→3) ⇒ GalNAc	C1GALT1, C1GALT1C1, B3GALT4, B3GALT5
Gal(β1→3) ⇒ Gal	B3GALT6
Gal(β1→3) ⇒ GlcNAc	B3GALT1, B3GALT2
Gal(α1→3) ⇒ Gal	ABO, GGTA1
Gal(β1→0) ⇒ Cer	UGT8
Gal(β1→4) ⇒ Xyl	B4GALT7
GalNAc(β1→4) ⇒ GlcA	GALNACT-2, CHSY1
GalNAc(α1→3) ⇒ GalNAc	GBGT1
GalNAc(α1→0) ⇒ Ser	GALNT1, GALNT2, GALNT3, GALNT4, GALNT5, GALNT6, GALNT7, GALNT8, GALNT9, GALNT10, GALNT11, GALNT12, GALNT13, GALNT14, GALNTL4
GalNAc(β1→4) ⇒ Gal	B4GALNT1, B4GALNT2
GalNAc(α1→0) ⇒ Thr	GALNT1, GALNT2, GALNT3, GALNT4, GALNT5, GALNT6, GALNT7, GALNT8, GALNT9, GALNT10, GALNT11, GALNT12, GALNT13, GALNT14, GALNTL4
GalNAc(α1→3) ⇒ Gal	ABO
GalNAc(β1→3) ⇒ Gal	B3GALNT1
GalNAc(β1→3) ⇒ GlcNAc	B3GALNT2
GalNAc(β1→4) ⇒ GlcNAc	B4GALNT3, B4GALNT4
Glc(β1→0) ⇒ Cer	UGCG, UGCGL1

Continued on next page

Table A.1 – continued from previous page

Reaction	**Genes**
Glc(α1→3) ⇒ Man	ALG6
Glc(α1→3) ⇒ Glc	ALG8
Glc(α1→2) ⇒ Glc	ALG10
Glc(β1→3) ⇒ Fuc	B3GALTL
GlcNAc(β1→3) ⇒ GalNAc	B3GNT6
GlcNAc(β1→4) ⇒ GlcNAc	ALG13, ALG14
GlcNAc(β1→0) ⇒ Ser	OGT
GlcNAc(β1→4) ⇒ GlcA	HAS1, HAS2, HAS3
GlcNAc(β1→0) ⇒ Thr	OGT
GlcNAc(β1→3) ⇒ Gal	B3GNT1, B3GNT2, B3GNT3, B3GNT4, B3GNT5, B3GNT6, B3GNTL1
GlcNAc(β1→2) ⇒ Man	POMGNT1, MGAT1, MGAT2
GlcNAc(β1→3) ⇒ Fuc	LFNG, RFNG
GlcNAc(β1→6) ⇒ GalNAc	GCNT1, GCNT3, GCNT4
GlcNAc(β1→6) ⇒ Man	MGAT5, MGAT9, MGAT5B
GlcNAc(α1→4) ⇒ Gal	A4GNT
GlcNAc(β1→6) ⇒ Gal	IGNT2, IGNT3, GCNT2, GCNT3
GlcNAc(α1→4) ⇒ GlcA	EXTL1, EXTL2, EXTL3
GlcNAc(β1→4) ⇒ Man	MGAT3, MGAT4A, MGAT4B, MGAT9
GlcA(β1→4) ⇒ GlcNAc	EXT1, EXT2
GlcA(β1→3) ⇒ GlcNAc	HAS1, HAS2, HAS3
GlcA(β1→3) ⇒ GalNAc	CHSY1
GlcA(β1→3) ⇒ Gal	B3GAT3, B3GAT1, B3GAT2
Man(α1→0) ⇒ Ser	POMT1, POMT2
Man(α1→3) ⇒ Man	ALG2, ALG3
Man(α1→0) ⇒ Thr	POMT1, POMT2
Man(β1→4) ⇒ GlcNAc	ALG1
Man(α1→6) ⇒ Man	ALG2, ALG12
Man(α1→4) ⇒ GlcN	PIGM
Man(α1→2) ⇒ Man	non-HomoSapiens, ALG9, PIGB
NeuAc(α2→6) ⇒ GalNAc	ST6GAL1, ST6GALNAC1, ST6GALNAC2, ST6GALNAC3, ST6GALNAC4, ST6GALNAC5, ST6GALNAC6
NeuAc(α2→8) ⇒ NeuAc	ST8SIA1, ST8SIA2, ST8SIA3, ST8SIA4, ST8SIA5, ST8SIA6

Continued on next page

Table A.1 – continued from previous page

Reaction	Genes
NeuAc(α2→6) ⇒ Gal	ST6GAL1, ST6GAL2, ST6GALNAC1, ST6GALNAC2, ST6GALNAC3, ST6GALNAC4
NeuAc(α2→3) ⇒ Gal	ST3GAL1, ST3GAL2, ST3GAL3, ST3GAL4, ST3GAL5, ST3GAL6
Xyl(β1→2) ⇒ Man	non-HomoSapiens
Xyl(β1→0) ⇒ Ser	XYLT1, XYLT2
Fuc(α1→3) ⇒ GlcNAc	FUT3, FUT4, FUT5, FUT6, FUT7, FUT11, FUT9
Fuc(α1→4) ⇒ GlcNAc	FUT3, FUT5, FUT6
Fuc(α1→0) ⇒ Ser	POFUT1, POFUT2
Fuc(α1→2) ⇒ Gal	FUT1, FUT2
Fuc(α1→6) ⇒ GlcNAc	FUT8
Fuc(α1→0) ⇒ Thr	POFUT1, POFUT2

A.2 Structures known to be artefacts

Glycosciences ID					
612	3994	11252	13424	16541	25759
624	3995	11258	14899	16542	26057
1172	3999	11259	15025	16598	
1241	4151	11260	15235	16685	
1737	4521	11261	15891	16732	
1774	4522	11352	16048	16735	
1775	4523	11732	16085	16736	
2652	4524	11733	16163	16737	
2740	4951	11734	16164	16850	
2747	4957	11735	16167	18548	
2854	5152	11736	16171	19743	
2921	5334	11785	16172	20279	
2922	5376	11786	16270	20282	
2924	6637	12157	16271	23725	
3152	6655	12209	16320	23726	
3697	6667	12210	16371	23737	
3823	7477	12211	16414	23738	
3825	7753	12212	16423	24339	
3826	7763	12213	16432	24340	
3827	8104	12445	16433	24341	
3846	8393	12833	16458	24403	
3847	8668	12867	16482	25613	
3848	9689	13305	16485	25678	
3849	10909	13306	16486	25679	
3961	10957	13322	16502	25699	

A.3 Human pathway diagrams

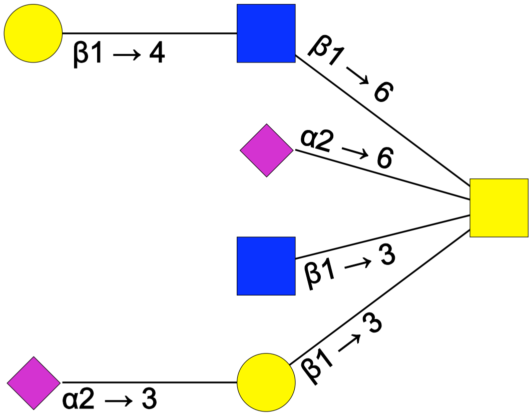


Figure A.1: Structure comprising structures along the O-linked pathway.

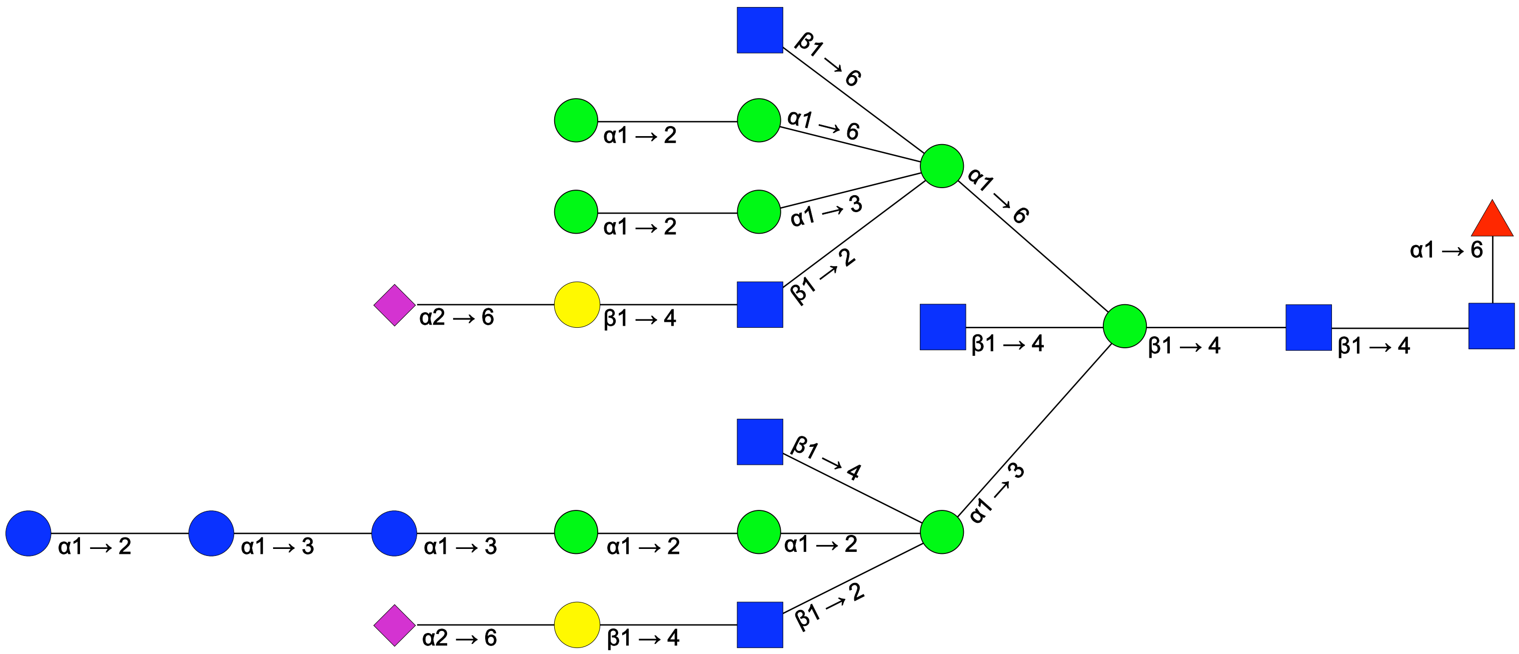


Figure A.2: Structure comprising structures along the N-linked pathway.

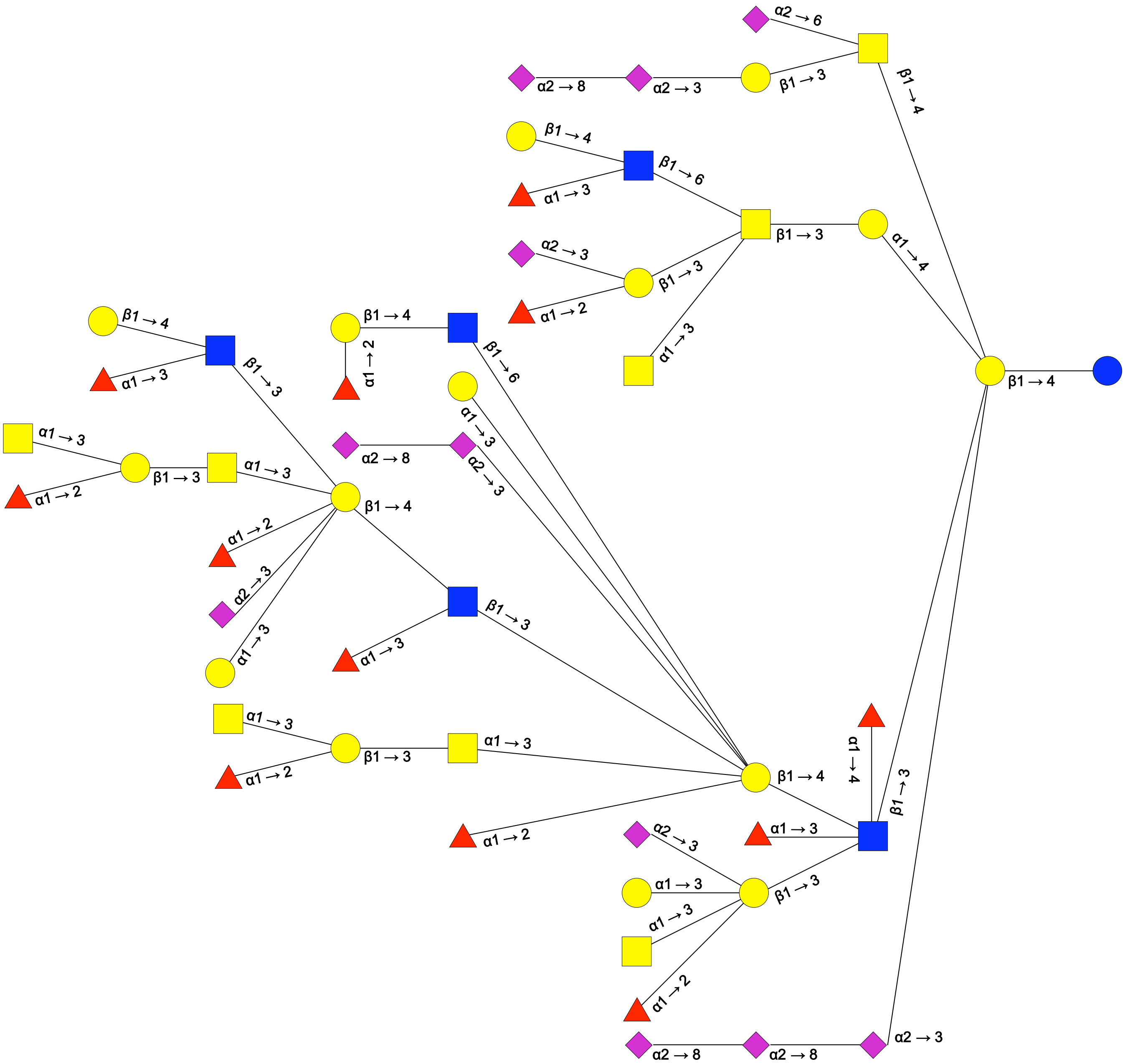


Figure A.3: Structure comprising structures along the GSL pathway.

BIBLIOGRAPHY

[1] Maureen E. Taylor and Kurt Drickamer. *Introduction to glycobiology*. Oxford University Press, Oxford ; New York, 2nd edition, 2006. 2005034575 GBA596525 Maureen E. Taylor, Kurt Drickamer. ill. ; 25 cm. Includes bibliographical references and index. Concepts of glycobiology – Conformations of oligosaccharides – N-linked glycosylation – O-linked glycosylation – Glycolipids and membrane protein glycosylation – Glycomics and analysis of glycan structures – Effects of glycosylation on protein structure and function – Carbohydrate recognition in cell adhesion and signalling – Glycoprotein trafficking in cells and organisms – Glycobiology of plants, bacteria, and viruses – Glycobiology and development – Glycosylation and disease – The future of glycobiology.

[2] S. Roseman. Reflections on glycobiology. *J Biol Chem*, 276(45):41527–42, 2001. GM51215/GM/United States NIGMS Journal Article Research Support, U.S. Gov't, P.H.S. Review United States. PMID:11553646.

[3] T. W. Rademacher, R. B. Parekh, and R. A. Dwek. Glycobiology. *Annu Rev Biochem*, 57:785–838, 1988. Journal Article Research Support, Non-U.S. Gov't Review United states. PMID:3052290.

[4] J. Stevens, O. Blixt, T. M. Tumpey, J. K. Taubenberger, J. C. Paulson, and I. A. Wilson. Structure and receptor specificity of the hemagglutinin from an h5n1 influenza virus. *Science*, 312(5772):404–10, 2006. AI058113/AI/United States NIAID AI42266/AI/United States NIAID CA55896/CA/United States NCI GM060938/GM/United States NIGMS GM062116/GM/United States NIGMS Journal Article Research Support, N.I.H., Extramural United States. PMID:16543414.

[5] A. Helenius and M. Aebi. Intracellular functions of n-linked glycans. *Science*, 291(5512):2364–9, 2001. Journal Article Research Support, Non-U.S. Gov't Review United States. PMID:11269317.

[6] S. Hengherr, A.G. Heyer, H.R. Köhler, and R.O. Schill. Trehalose and anhydrobiosis in tardigrades–evidence for divergence in responses to dehydration. *FEBS J.*, 275:281–288, Jan 2008. PMID:18070104.

[7] J. B. Lowe. Glycan-dependent leukocyte adhesion and recruitment in inflammation. *Curr Opin Cell Biol*, 15(5):531–8, 2003. 1P01CA71932/CA/United States NCI GM 62116/GM/United States NIGMS Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review United States. PMID:14519387.

[8] K. S. Lau, E. A. Partridge, A. Grigorian, C. I. Silvescu, V. N. Reinhold, M. Demetriou, and J. W. Dennis. Complex n-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. *Cell*, 129(1):123–34, 2007. Journal Article Research Support, Non-U.S. Gov't United States. PMID:17418791.

[9] J. B. Lowe and J. D. Marth. A genetic approach to mammalian glycan function. *Annu Rev Biochem*, 72:643–91, 2003. 1P01-CA71932/CA/United States NCI DK48247/DK/United States NIDDK GM62116/GM/United States NIGMS P01HL-57345/HL/United States NHLBI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review United States. PMID:12676797.

[10] H. H. Freeze. Genetic defects in the human glycome. *Nat Rev Genet*, 7(7):537–51, 2006. Journal Article Research Support, N.I.H., Extramural Review England. PMID:16755287.

[11] K. Ohtsubo and J. D. Marth. Glycosylation in cellular mechanisms of health and disease. *Cell*, 126(5):855–67, 2006. DK4247/DK/United States NIDDK GM62116/GM/United States NIGMS HL57345/HL/United States NHLBI Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review United States. PMID:16959566.

[12] L. A. Tabak. In defense of the oral cavity: structure, biosynthesis, and function of salivary mucins. *Annu Rev Physiol*, 57:547–64, 1995. Journal Article Research Support, U.S. Gov't, P.H.S. Review United states. PMID:7778877.

[13] T. Feizi and B. Mulloy. Carbohydrates and glycoconjugates. glycomics: the new era of carbohydrate biology. *Curr Opin Struct Biol*, 13(5):602–4, 2003. Editorial England. PMID:14568615.

[14] A. Dell and H. R. Morris. Glycoprotein structure determination by mass spectrometry. *Science*, 291(5512):2351–6, 2001. Journal Article Review United States. PMID:11269315.

[15] P. M. Rudd, C. Colominas, L. Royle, N. Murphy, E. Hart, A. H. Merry, H. F. Hebestreit, and R. A. Dwek. A high-performance liquid chromatography based strategy for rapid, sensitive sequencing of n-linked oligosaccharide modifications to proteins in sodium dodecyl sulphate polyacrylamide electrophoresis gel bands. *Proteomics*, 1(2):285–94, 2001. Journal Article Germany. PMID:11680875.

[16] J. Duus, C. H. Gotfredsen, and K. Bock. Carbohydrate structural determination by nmr spectroscopy: modern methods and limitations. *Chem Rev*, 100(12):4589–614, 2000. Journal Article United States. PMID:11749359.

[17] H. Tang, Y. Mechref, and M. V. Novotny. Automated interpretation of ms/ms spectra of oligosaccharides. *Bioinformatics*, 21 Suppl 1:i431–i439, 2005. 1367-4803 Journal Article. PMID:15961488.

[18] H. J. Joshi, M. J. Harrison, B. L. Schulz, C. A. Cooper, N. H. Packer, and N. G. Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics*, 4(6):1650–64, 2004. Journal Article Germany. PMID:15174134.

[19] K. Maass, R. Ranzinger, H. Geyer, C. W. von der Lieth, and R. Geyer. Glyco-peakfinder–de novo composition analysis of glycoconjugates. *Proteomics*, 7(24):4435–44, 2007. Journal Article Research Support, Non-U.S. Gov't Germany. PMID:18072204.

[20] K. K. Lohmann and C. W. von der Lieth. Glyco-fragment: A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics*, 3(10):2028–35, 2003. Journal Article Research Support, U.S. Gov't, Non-P.H.S. Germany. PMID:14625865.

[21] K. K. Lohmann and C. W. von der Lieth. Glycofragment and glycosearchms: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res*, 32(Web Server issue):W261–6, 2004. Journal Article England. PMID:15215392.

[22] Bertram Fraser-Reid. Glycoscience : chemistry and chemical biology. pages 2219–40. Springer, New York, 2nd edition, 2008. 2008921863 Bertram Fraser-Reid.

[23] E. M. Comelli, M. Amado, S. R. Head, and J. C. Paulson. Custom microarray for glycobiologists: considerations for glycosyltransferase gene expression profiling. *Biochem Soc Symp*, (69):135–42, 2002. Journal Article Review England. PMID:12655780.

[24] W. Kemmner, C. Roefzaad, W. Haensch, and P. M. Schlag. Glycosyltransferase expression in human colonic tissue examined by oligonucleotide arrays. *Biochim Biophys Acta*, 1621(3):272–9, 2003. Evaluation Studies Journal Article Netherlands. PMID:12787925.

[25] T. Feizi, F. Fazio, W. Chai, and C. H. Wong. Carbohydrate microarrays - a new set of technologies at the frontiers of glycomics. *Curr Opin Struct Biol*, 13(5):637–45, 2003. Journal Article Research Support, Non-U.S. Gov't Review England. PMID:14568620.

[26] C. W. von der Lieth, A. Bohne-Lang, K. K. Lohmann, and M. Frank. Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform*, 5(2):164–78, 2004. 1467-5463 Journal Article. PMID:15260896.

[27] C. W. von der Lieth, T. Lutteke, and M. Frank. The role of informatics in glycobiology research with special emphasis on automatic interpretation of ms spectra. *Biochim Biophys Acta*, 1760(4):568–77, 2006. Journal Article Research Support, Non-U.S. Gov't Review Netherlands. PMID:16459020.

[28] K. F. Aoki-Kinoshita. An introduction to bioinformatics for glycomics research. *PLoS Comput Biol*, 4(5):e1000075, 2008. Journal Article Review United States. PMID:18516240.

[29] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. The universal protein resource (uniprot). *Nucleic Acids Res*, 33(Database issue):D154–9, 2005. 1R01HGO2273-01/HG/United States NHGRI U01 HG02712-01/HG/United States NHGRI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. England. PMID:15608167.

[30] K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. Kegg as a glycome informatics resource. *Glycobiology*, 16(5):63R–70R, 2006. Journal Article Research Support, Non-U.S. Gov't Review England. PMID:16014746.

[31] T. Lutteke, A. Bohne-Lang, A. Loss, T. Goetz, M. Frank, and C. W. von der Lieth. Glycosciences.de: an internet portal to support glycomics and glycobiology research. *Glycobiology*, 16(5):71R–81R, 2006. Journal Article Research Support, Non-U.S. Gov't Review England. PMID:16239495.

[32] F.V Toukach and Knirel Y. New database of bacterial carbohydrate structures. In *XVIII International Symposium on Glycoconjugates*, pages 216–217, Florence, 2005.

[33] Glycominds website (http://www.glycominds.com) [online, cited 18/05/07].

[34] S. Doubet, K. Bock, D. Smith, A. Darvill, and P. Albersheim. The complex carbohydrate structure database. *Trends Biochem Sci*, 14(12):475–7, 1989. Journal Article England. PMID: 2623761.

[35] Y. Hizukuri, Y. Yamanishi, K. Hashimoto, and M. Kanehisa. Extraction of species-specific glycan substructures. *Genome Inform Ser Workshop Genome Inform*, 15(1):69–81, 2004. 0919-9454 (Print) Journal Article. PMID:15712111.

[36] Y. Hizukuri, Y. Yamanishi, O. Nakamura, F. Yagi, S. Goto, and M. Kanehisa. Extraction of leukemia specific glycan motifs in humans by computational glycomics. *Carbohydr Res*, 340(14):2270–8, 2005. Journal Article Netherlands. PMID:16095580.

[37] C. A. Cooper, H. J. Joshi, M. J. Harrison, M. R. Wilkins, and N. H. Packer. Glycosuitedb: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res*, 31(1):511–3, 2003. Journal Article England. PMID:12520065.

[38] Wikipedia. Wikipedia — Wikipedia, the free encyclopedia, 2008. [Online; accessed 23-July-2008].

[39] S. Kawano, K. Hashimoto, T. Miyama, S. Goto, and M. Kanehisa. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics*, 21(21):3976–82, 2005. 1367-4803 (Print) Journal Article. PMID:16159923.

[40] K. Drickamer and M. E. Taylor. Glycan arrays for functional glycomics. *Genome Biol*, 3(12):REVIEWS1034, 2002. Journal Article Review England. PMID:12537579.

[41] E. M. Comelli, S. R. Head, T. Gilmartin, T. Whisenant, S. M. Haslam, S. J. North, N. K. Wong, T. Kudo, H. Narimatsu, J. D. Esko, K. Drickamer, A. Dell, and J. C. Paulson. A focused microarray approach to functional glycomics: transcriptional regulation of the glycome. *Glycobiology*, 16(2):117–31, 2006. AI50143/AI/United States NIAID GM33063/GM/United States NIGMS GM62116/GM/United States NIGMS United Kingdom Wellcome Trust Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England. PMID:16237199.

[42] M. Yamamoto, F. Yamamoto, T. T. Luong, T. Williams, Y. Kominato, and F. Yamamoto. Expression profiling of 68 glycosyltransferase genes in 27 different human tissues by the systematic multiplex reverse transcription-polymerase chain reaction method revealed clustering of sexually related tissues in hierarchical clustering algorithm analysis. *Electrophoresis*, 24(14):2295–307, 2003. 0173-0835 (Print) Journal Article. PMID:12874863.

[43] Alison V. Nairn, William S. York, Kyle Harris, Erica M. Hall, J. Michael Pierce, and Kelley W. Moremen. Regulation of glycan structures in animal tissues: Transcript profiling of glycan-related genes. *J. Biol. Chem.*, page M801964200, 2008.

[44] V. Malhotra and S. Mayor. Cell biology: the golgi grows up. *Nature*, 441(7096):939–40, 2006. Comment News England. PMID:16791181.

[45] R. J. Ivatt. Regulation of glycoprotein biosynthesis by formation of specific glycosyltransferase complexes. *Proc Natl Acad Sci U S A*, 78(7):4021–5, 1981. CA14051/CA/United States NCI CA14142/CA/United States NCI Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United states. PMID:6457298.

[46] C. G. Giraudo and H. J. Maccioni. Ganglioside glycosyltransferases organize in distinct multienzyme complexes in cho-k1 cells. *J Biol Chem*, 278(41):40262–71, 2003. Journal Article Research Support, Non-U.S. Gov't United States. PMID:12900410.

[47] C. L. de Graffenried and C. R. Bertozzi. The roles of enzyme localisation and complex formation in glycan assembly within the golgi apparatus. *Curr Opin Cell Biol*, 16(4):356–63, 2004. GM59907/GM/United States NIGMS Journal Article Research Support, U.S. Gov't, P.H.S. Review United States. PMID:15261667.

[48] K. Y. Do, N. Fregien, M. Pierce, and R. D. Cummings. Modification of glycoproteins by n-acetylglucosaminyltransferase v is greatly influenced by accessibility of the enzyme to oligosaccharide acceptors. *J Biol Chem*, 269(38):23456–64, 1994. CA-37626/CA/United States NCI Journal Article Research Support, U.S. Gov't, P.H.S. United states. PMID:7522229.

[49] J. U. Baenziger. Protein-specific glycosyltransferases: how and why they do it! *Faseb J*, 8(13):1019–25, 1994. CA-21923/CA/United States NCI DK-41738/DK/United States NIDDK Journal Article Research Support, U.S. Gov't, P.H.S. Review United states official publication of the Federation of American Societies for Experimental Biology. PMID:7926366.

[50] J. J. Caramelo and A. J. Parodi. How sugars convey information on protein conformation in the endoplasmic reticulum. *Semin Cell Dev Biol*, 18(6):732–42, 2007. Journal Article Review England. PMID:17997334.

[51] P. Stanley. Glycosylation mutants of animal cells. *Annu Rev Genet*, 18:525–52, 1984. P30 CA13330-12/CA/United States NCI R01 CA30645/CA/United States NCI R01 CA36434/CA/United States NCI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review United states. PMID:6241454.

[52] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson, and R. Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods*, 2(11):817–24, 2005. U54 GM62116/GM/United States NIGMS Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, P.H.S. Review United States. PMID:16278650.

[53] F. J. Krambeck and M. J. Betenbaugh. A mathematical model of n-linked glycosylation. *Biotechnol Bioeng*, 92(6):711–28, 2005. 0006-3592 (Print) Journal Article. PMID:16247773.

[54] P. Hossler, B. C. Mulukutla, and W. S. Hu. Systems analysis of n-glycan processing in mammalian cells. *PLoS ONE*, 2(1):e713, 2007. Journal Article United States. PMID:17684559.

[55] J. A. Campbell, G. J. Davies, V. Bulone, and B. Henrissat. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J*, 326 (Pt 3):929–39, 1997. Letter Research Support, Non-U.S. Gov't England. PMID:9334165.

[56] P. M. Coutinho, E. Deleury, G. J. Davies, and B. Henrissat. An evolving hierarchical family classification for glycosyltransferases. *J Mol Biol*, 328(2):307–17, 2003. Journal Article Research Support, Non-U.S. Gov't England. PMID:12691742.

[57] R. Ranzinger and S. Herget. Glycome db [online]. Date 2008.

[58] H. Narimatsu. Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J*, 21(1-2):17–24, 2004. Journal Article Research Support, Non-U.S. Gov't Review United States. PMID:15467393.

[59] A. D. McNaught. International union of pure and applied chemistry and international union of biochemistry and molecular biology. joint commission on biochemical nomenclature. nomenclature of carbohydrates. *Carbohydr Res*, 297(1):1–92, 1997. Journal Article Netherlands. PMID:9042704.

[60] S. Herget, R. Ranzinger, K. Maass, and C. W. Lieth. Glycoct-a unifying sequence format for carbohydrates. *Carbohydr Res*, 2008. Journal article. PMID:18436199.

[61] S. S. Sahoo, C. Thomas, A. Sheth, C. Henson, and W. S. York. Glyde-an expressive xml standard for the representation of glycan structure. *Carbohydr Res*, 340(18):2802–7, 2005. 5 P41 RR18502-02/RR/United States NCRR Journal Article Research Support, N.I.H., Extramural Netherlands. PMID:16242678.

[62] David Flanagan and Yukihiro Matsumoto. *The Ruby programming language*. O'Reilly, Beijing ; Sebastopol, CA, 1st edition, 2008. 2008297501 David Flanagan and Yukihiro Matsumoto. ill. ; 24 cm. Covers Ruby 1.8 and 1.9–Cover.

[63] J. C. Michalski, J. Lemoine, J. M. Wieruszeski, B. Fournet, J. Montreuil, and G. Strecker. Characterization of a novel type of chain-terminator gal beta 1-6gal beta 1-4)glcnac in an oligosaccharide related to n-glycosylated protein glycans isolated from gm1 the urine of patients with gangliosidosis. *Eur J Biochem*, 198(2):521–6, 1991. Journal Article Research Support, Non-U.S. Gov't Germany. PMID:1904026.

[64] R. J. Pollitt and K. M. Pretty. The glycoasparagines in urine of a patient with aspartylglycosaminuria. *Biochem J*, 141(1):141–6, 1974. Journal Article England. PMID:4455197.

[65] J. Angstrom, H. Karlsson, K. A. Karlsson, G. Larson, and K. Nilson. Galnac beta 1—-3 terminated glycosphingolipids of human erythrocytes. *Arch Biochem Biophys*, 251(2):440–9, 1986. Journal Article Research Support, Non-U.S. Gov't United states. PMID:3800377.

[66] J. Kumlien, G. Gronberg, B. Nilsson, O. Mansson, D. Zopf, and A. Lundblad. Structural and immunochemical analysis of three alpha-limit dextrin oligosaccharides. *Arch Biochem Bio-*

phys, 269(2):678–89, 1989. Journal Article Research Support, Non-U.S. Gov't United states. PMID:2919890.

[67] W. P. Aston, A. S. Donald, and W. T. Morgan. A trisaccharide, o-beta-d-galactopyranosyl-(1–¿3)-o-(n- acetyl-beta-d-glucosaminopyranosyl)-(1–¿4)-d- galactose, obtained from human blood-group h substance. *Biochem J*, 107(6):861–863, 1968. Journal article. PMID:16742612.

[68] S. K. Kundu, Y. Harati, and L. K. Misra. Sialosylglobotetraosylceramide: a marker for amyotropic lateral sclerosis. *Biochem Biophys Res Commun*, 118(1):82–9, 1984. Comparative Study Journal Article United states. PMID:6696769.

[69] C. C. Sung, D. K. Pearl, S. W. Coons, B. W. Scheithauer, P. C. Johnson, and A. J. Yates. Gangliosides as diagnostic markers of human astrocytomas and primitive neuroectodermal tumors. *Cancer*, 74(11):3010–22, 1994. CA 50910/CA/United States NCI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states. PMID:7954264.

[70] P. H. Johnson, A. S. Donald, J. Feeney, and W. M. Watkins. Reassessment of the acceptor specificity and general properties of the lewis blood-group gene associated alpha-3/4-fucosyltransferase purified from human milk. *Glycoconj J*, 9(5):251–64, 1992. Comparative Study Journal Article England. PMID:1490104.

[71] K. Yamashita, Y. Tachibana, and A. Kobata. Oligosaccharides of human milk. isolation and characterization of three new disialyfucosyl hexasaccharides. *Arch Biochem Biophys*, 174(2):582–91, 1976. Journal Article United states. PMID:1230009.

[72] D. F. Smith, D. A. Zopf, and V. Ginsburg. Fractionation of sialyl oligosaccharides of human milk by ion-exchange chromatography. *Anal Biochem*, 85(2):602–8, 1978. Journal Article United states. PMID:646116.

[73] G. Strecker, A. Pierce-Cretel, B. Fournet, G. Spik, and J. Montreuil. Characterization by gas–liquid chromatography–mass spectrometry of oligosaccharides resulting from the hydrazinolysis-nitrous acid deamination reaction of glycopeptides. *Anal Biochem*, 111(1):17–26, 1981. Journal Article Research Support, Non-U.S. Gov't United states. PMID:7235235.

[74] A. Leppanen, A. Korvuo, K. Puro, and O. Renkonen. Glycoproteins of human teratocarcinoma cells (pa1) carry both anomers of o-glycosyl-linked d-galactopyranosyl-(1—-3)-2-acetamido- 2-deoxy-alpha-d-galactopyranosyl group. *Carbohydr Res*, 153(1):87–95, 1986. Journal Article Research Support, Non-U.S. Gov't Netherlands. PMID:3779692.

[75] V. K. Dua, B. N. Rao, S. S. Wu, V. E. Dube, and C. A. Bush. Characterization of the oligosaccharide alditols from ovarian cyst mucin glycoproteins of blood group a using high pressure

liquid chromatography (hplc) and high field 1h nmr spectroscopy. *J Biol Chem*, 261(4):1599–608, 1986. GM 31449/GM/United States NIGMS Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United states. PMID:3003076.

[76] H. van Halbeek, A. M. Strang, M. Lhermitte, H. Rahmoune, G. Lamblin, and P. Roussel. Structures of monosialyl oligosaccharides isolated from the respiratory mucins of a non-secretor (o, lea+b-) patient suffering from chronic bronchitis. characterization of a novel type of mucin carbohydrate core structure. *Glycobiology*, 4(2):203–19, 1994. P41-RR-05351/RR/United States NCRR R01-HL-38213/HL/United States NHLBI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. England. PMID:8054719.

[77] J. C. Michalski, S. Bouquelet, J. Montreuil, G. Strecker, O. Dulac, and A. Munnich. Abnormal galactoside excretion in urine of a patient with early myoclonic epileptic encephalopathy. *Clin Chim Acta*, 137(1):43–51, 1984. Journal Article Research Support, Non-U.S. Gov't Netherlands. PMID:6421512.

[78] M. Messer, E. Trifonoff, W. Stern, J. G. Collins, and J. H. Bradbury. Structure of a marsupial-mild trisaccharide. *Carbohydr Res*, 83(2):327–34, 1980. Comparative Study Journal Article Netherlands. PMID:7407802.

[79] T. Hiruma, A. Togayachi, K. Okamura, T. Sato, N. Kikuchi, Y. D. Kwon, A. Nakamura, K. Fujimura, M. Gotoh, K. Tachibana, Y. Ishizuka, T. Noce, H. Nakanishi, and H. Narimatsu. A novel human beta1,3-n-acetylgalactosaminyltransferase that synthesizes a unique carbohydrate structure, galnacbeta1-3glcnac. *J Biol Chem*, 279(14):14087–95, 2004. Journal Article Research Support, Non-U.S. Gov't United States. PMID:14724282.

[80] F. Piller, D. Blanchard, M. Huet, and J. P. Cartron. Identification of a alpha-neuac-(2—-3)-beta-d-galactopyranosyl n-acetyl-beta-d-galactosaminyltransferase in human kidney. *Carbohydr Res*, 149(1):171–84, 1986. Journal Article Research Support, Non-U.S. Gov't Netherlands. PMID:2425965.

[81] T. Endo, J. Amano, E. G. Berger, and A. Kobata. Structure identification of the complex-type, asparagine-linked sugar chains of beta-d-galactosyl-transferase purified from human milk. *Carbohydr Res*, 150:241–63, 1986. Journal Article Research Support, Non-U.S. Gov't Netherlands. PMID:3093076.

[82] P. Hermentin, R. Witzel, R. Doenges, R. Bauer, H. Haupt, T. Patel, R. B. Parekh, and D. Brazel. The mapping by high-ph anion-exchange chromatography with pulsed amperometric detection and capillary electrophoresis of the carbohydrate moieties of human plasma alpha 1-acid glycoprotein. *Anal Biochem*, 206(2):419–29, 1992. Journal Article United states. PMID:1443615.

[83] A. Mizoguchi, T. Mizuochi, and A. Kobata. Structures of the carbohydrate moieties of secretory component purified from human milk. *J Biol Chem*, 257(16):9612–21, 1982. Journal Article Research Support, Non-U.S. Gov't United states. PMID:7107583.

[84] D. K. Podolsky. Oligosaccharide structures of human colonic mucin. *J Biol Chem*, 260(14):8262–71, 1985. AM01257/AM/United States NIADDK AM34422/AM/United States NIADDK Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states. PMID:4008490.

[85] M. R. Stroud, S. B. Levery, E. D. Nudelman, M. E. Salyan, J. A. Towell, C. E. Roberts, M. Watanabe, and S. Hakomori. Extended type 1 chain glycosphingolipids: dimeric lea (iii4v4fuc2lc6) as human tumor-associated antigen. *J Biol Chem*, 266(13):8439–46, 1991. CA42505/CA/United States NCI Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states. PMID:2022659.

[86] T. Yagyu, T. Monden, M. Baba, Y. Tamaki, T. Takeda, T. Kobayashi, T. Shimano, Y. Tsuji, H. Matsushita, H. Osawa, and et al. A cancer-reactive human monoclonal antibody derived from a colonic cancer patient treated with local immunotherapy. *Jpn J Cancer Res*, 84(1):75–82, 1993. Journal Article Research Support, Non-U.S. Gov't Japan Gann. PMID:8449830.

[87] J. Beuth, H. L. Ko, G. Pulverer, G. Uhlenbruck, and H. Pichlmaier. Importance of lectins for the prevention of bacterial infections and cancer metastases. *Glycoconj J*, 12(1):1–6, 1995. Journal Article Review England. PMID:7795408.

[88] R. J. Harris, H. van Halbeek, J. Glushka, L. J. Basa, V. T. Ling, K. J. Smith, and M. W. Spellman. Identification and structural analysis of the tetrasaccharide neuac alpha(2–¿6)gal beta(1–¿4)glcnac beta(1–¿3)fuc alpha 1–¿o-linked to serine 61 of human factor ix. *Biochemistry*, 32(26):6539–47, 1993. P41-RR-05351/RR/United States NCRR Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states. PMID:8329384.

[89] N. Kuraya, K. Omichi, H. Nishimura, S. Iwanaga, and S. Hase. Structural analysis of o-linked sugar chains in human blood clotting factor ix. *J Biochem*, 114(6):763–5, 1993. Journal Article Research Support, Non-U.S. Gov't Japan. PMID:8138528.

[90] M. von Der Ohe, S. F. Wheeler, M. Wuhrer, D. J. Harvey, S. Liedtke, M. Muhlenhoff, R. Gerardy-Schahn, H. Geyer, R. A. Dwek, R. Geyer, D. R. Wing, and M. Schachner. Localization and characterization of polysialic acid-containing n-linked glycans from bovine ncam. *Glycobiology*, 12(1):47–63, 2002. Journal Article Research Support, Non-U.S. Gov't England. PMID:11825886.

[91] R. Kannagi, S. B. Levery, and S. Hakomori. Lea-active heptaglycosylceramide, a hybrid of type 1 and type 2 chain, and the pattern of glycolipids with lea, leb, x (lex), and y (ley) deter-

minants in human blood cell membranes (ghosts). evidence that type 2 chain can elongate repetitively but type 1 chain cannot. *J Biol Chem*, 260(10):6410–5, 1985. CA19224/CA/United States NCI CA20026/CA/United States NCI Comparative Study Journal Article Research Support, U.S. Gov't, P.H.S. United states. PMID:3997830.

[92] F. G. Hanisch, G. Uhlenbruck, J. Peter-Katalinic, H. Egge, J. Dabrowski, and U. Dabrowski. Structures of neutral o-linked polylactosaminoglycans on human skim milk mucins. a novel type of linearly extended poly-n-acetyllactosamine backbones with gal beta(1-4)glcnac beta(1-6) repeating units. *J Biol Chem*, 264(2):872–83, 1989. Journal Article Research Support, Non-U.S. Gov't United states. PMID:2910868.

[93] J Gosling, B Joy, and GL Steele. *The Java Language Specification*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 1996.

[94] Tranche website (http://tranche.proteomecommons.org/) [online].

[95] H. J. Joshi. Implementation of p2p-network and central database. Technical report, DKFZ, 2007.

[96] P. Toukach, H. J. Joshi, R. Ranzinger, Y. Knirel, and C. W. von der Lieth. Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the bacterial carbohydrate structure database and glycosciences.de. *Nucleic Acids Res*, 35(Database issue):D280–6, 2007. Journal Article Research Support, Non-U.S. Gov't England. PMID:17202164.

[97] Hibernate website (http://www.hibernate.org) [online].

[98] Struts website (http://struts.apache.org/2.x/) [online].

[99] Tomcat website (http://tomcat.apache.org/) [online].

[100] Postgresql website (http://www.postgresql.org/) [online].

[101] Jibx sourceforge (http://jibx.sourceforge.net) [online].

[102] Xfire sourceforge (http://xfire.sourceforge.net) [online].

[103] Doug Zongker. Chicken chicken chicken: Chicken chicken. pages 16–21.

Acknowledgements

The all important acknowledgements section, a chance to formally recognise that I could not have completed this project and thesis without the help of so many other people.

First and foremost, I must thank my supervisor — Willi von der Lieth, who died in November of 2007. It's due to his support that I not only started doing this PhD, but managed to complete it successfully. His influence in the field is already sorely missed, and one can only hope that we can continue working on his legacy.

I must also give thanks to the puppet masters holding the purse strings from the EUROCarbDB project and DKFZ. Thanks for giving me enough money for me to engage in the mental marathon that is the PhD. To the European union — god bless your liberal funding policies.

A word to my colleagues, both in my lab, and from the EUROCarbDB project in general. Thanks for the reality checks — they helped me to remain grounded. To my ECDB colleagues in particular, thank-you for teaching me the true meaning of meetings — may there be many more Porterhouse projects in the future. Special thanks to Alessio for taking the time to read my thesis.

To Caroline I give my heartfelt thanks for transforming what I wrote into English. I still can't believe how good a job you did in fixing up my thesis. You're truly a professional. Just ignore the errors in this acknowledgements section.

To my family — who had originally convinced me to do a PhD thesis. I hope you're happy now Mum. What can I write to give justice to a lifetime of support from my parents? I think I did you guys proud.

I have to give special thanks to my friends here, the people who have acted as my steadfast support throughout this whole process. To everyone I've had beers with, "Prost!". To Nuria, Anita, Aaron, Maria and especially Birgit — I am unbelievably appreciative of you guys. Whether it's making pancakes at 5am, running under fireworks, dragging my sorry and broken self in to a hospital, or knowing me well enough to put me back on the right track — I thank you all. They say the PhD is all about staying sane [103] long enough to hand it in, thanks for carrying me over the line.

Finally, I'd like to apologise to Sophia. I promise I'll make it up to you in the future.