12-2011

DATA-INTENSIVE COMPUTING FOR BIOINFORMATICS USING VIRTUALIZATION TECHNOLOGIES AND HPC INFRASTRUCTURES

Pengfei Xuan
Clemson University, pfxuan@gmail.com

DATA-INTENSIVE COMPUTING FOR BIOINFORMATICS USING VIRTUALIZATION TECHNOLOGIES AND HPC INFRASTRUCTURES

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
School of Computing

---

by
Pengfei Xuan
December 2011

Accepted by:
Dr. Feng Luo, Committee Chair
Dr. Amy W. Apon
Dr. Anna V. Blenda

ABSTRACT

The bioinformatics applications often involve many computational components and massive data sets, which are very difficult to be deployed on a single computing machine. In this thesis, we designed a data-intensive computing platform for bioinformatics applications using virtualization technologies and high performance computing (HPC) infrastructures with the concept of multi-tier architecture, which can seamlessly integrate the web user interface (presentation tier), scientific workflow (logic tier) and computing infrastructure (data/computing tier). We demonstrated our platform on two bioinformatics projects. First, we redesigned and deployed the cotton marker database (CMD) (http://www.cottonmarker.org), a centralized web portal in the cotton research community, using the Xen-based virtualization solution. To achieve high-performance and scalability for CMD web tools, we hosted the large amounts of protein databases and computational intensive applications of CMD on the Palmetto HPC of Clemson University. Biologists can easily utilize both bioinformatics applications and HPC resources through the CMD website without a background in computer science. Second, we developed a web tools ─ Glycan Array QSAR Tool (http://bci.clemson.edu/tools/glycan_array), to analyze glycan array data. The user interface of this tool was developed at the top of Drupal Content Management Systems (CMS) and the computational part was implemented using MATLAB Compiler Runtime (MCR) module. Our new bioinformatics computing platform enables the rapid deployment of data-intensive bioinformatics applications on HPC and virtualization environment with a user-friendly web interface and bridges the gap between biological scientists and cyberinfrastructure.

ACKNOWLEDGMENTS

Foremost, I would like to thank Dr. Anna Blenda, the Principal Investigator of the Cotton Marker Database development (project 03-391 funded by the Cotton Incorporated), for the financial support of my graduate research assistantship and the opportunity to do my research using the CMD resources. In addition, I would like to thank Dr. Blenda for her advice and friendship. One of the primary benefits of working for Dr Blenda is the freedom she has given me to design my own solutions to the CMD project and perform the further exploration of the solutions.

I would like to express my sincere gratitude to my advisor, Dr. Feng Luo for the continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of the thesis and papers. His ability to answer fundamental questions about complex problems without losing sight of the grand scheme is an ability that I hope to someday acquire. Dr. Luo has played a large and unique part in helping me reach this point in my journey.

I would also like to expressly thank Dr. Amy Apon for being on my committee, reviewing my dissertation and her insight into how to continue my future research.

I am very grateful to the remaining members: Dr. Bo Li, Yuehua Zhang, Aditya Sriram and Gauri Jape (BCI research group), Dr. Stephen Ficklin, Dr. Meg Staton and Dr. Chris Saski (CUGI laboratory), Dr. Véronique Decroocq (INRA, France) and Dr. Jean-Marc Lacape (UMR-DAP, France), for their research collaboration, support, and friendship.

Finally, I would like to thank my wife – Yueli Zheng, my daughter – Emma Xuan, my parents Yuhai Xuan and Jingzhen Li, and my mother-in-law Ailian Zhao. I would not have achieved this without their constant love and support.

TABLE OF CONTENTS

Page

Table of Contents (Continued)

Page

Table of Contents (Continued)

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

List of Figures (Continued)

Figure Page

ACRONYM TABLE

AUC	Area Under Curve
BLAST	Basic Local Alignment Search Tool
BCI	Bioinformatics and Chemical Informatics
CFG	Consortium for Functional Glycomics
CMD	Cotton Marker Database
CMS	Content Management System
EST	Expressed Sequence Tags
GMOD	Generic Model Organism Database
LVM	Logical Volume Manager
LV	Logical Volume
MCR	MATLAB Compiler Runtime
MPI	Message Passing Interface
MX	Mail Exchange
NGS	Next Generation Sequencing
ORFs	Open Reading Frames
PBS	Portable Batch System
PLS	Partial Least Squares
QSAR	Quantitative Structure-activity Relationship
QTL	Quantitative Trait Locus
ROC	Receiver Operating Characteristic
SNP	Single-nucleotide Polymorphism
SSR	Simple Sequence Repeats
SVM	Support Vector Machines
VM	Virtual Machine

Chapter 1

Introduction

The bioinformatics application often involves many computational components and a large number of runs with different parameters and configurations [1-4]. Recently, the genomic research based on Next Generation Sequencing (NGS) technology makes it possible to study biological phenomena on a large scale: all metabolic processes in a tissue, all transcripts in a cell, and all genes in a genome. However, sequencing-based genome-wide analysis also products massive quantities of data, which brings a new challenge facing the current bioinformatics research [5] and finally leads to the insufficiency of performance or capability on a local workstation or server. For example, the amount of data from the 1000 Genomes Project [6] will reach the petabyte scale for the raw sequence information. In the near future, the situation will be dramatically changed by third-generation sequencing technologies [7]. It will allow us to scan entire transcriptomes, microbiomes and genomes, and make it possible to assess epigenetic changes directly [8] in a few minutes with the cost less than US$100. Today's data-intensive biology drives a new emerging computational model to integrate massive data with computing resources derived from molecular biology.

Biological research is becoming more and more dependent on big data analysis. Biologists will soon encounter difficulties in handling missive data sets using traditional applications or tools that were initially designed for the single machine or non-data-intensive task in such areas as protein or nucleic acid sequence assembly, sequence

alignment for similarity comparisons, motif recognition in linear sequences or higher-order structure, and common patterns of gene expression; Further, while current technologies can provide large-scale computing solutions to analyze, integrate and manipulate big data sets, there still exists a huge knowledge gap for biologists with sufficient computer science background to utilize high-end platform resources. This shortage means biologists have to either avoid research areas relevant to big data sets or collaborate with data scientists. Due to lack of the ability to analyze the massive data, many promising biological projects have to be given up or cannot keep growing in long-term investments, and the ecosystem of bioinformatics research would also be limited to a very small scope.

High-performance Computing (HPC) is a typical computing infrastructure that provides high-performance parallel file systems with both high bandwidth and large capacity. Parallel file systems like PVFS [9], HDFS [10], GPFS [11] and Lustre [12] are usually employed to meet these needs and are often layered on top of clustered storage systems or use special high-end customized hardware. These systems are geared towards storing petabyte-scale data in a reliable fashion with high throughput for massive computing jobs. To make the bioinformatics resource and software more accessible, and allow for a faster overall time-to-solution, a user-friendly computing platform based on cyberinfrastructure could be a promising way to lead biologists to a new scientific paradigm: data-driven science.

In this thesis, we propose a platform solution for data-intensive bioinformatics applications using virtualization technologies and HPC infrastructures. A multi-tier

architecture is used to integrate the web user interface (presentation tier), scientific workflow (logic tier) and computing infrastructure (data/computing tier). The platform has been demonstrated through two bioinformatics projects: One is Cotton Marker Database (CMD) (http://www.cottonmarker.org) which is a centralized web portal in the cotton research community. The other project is a quantitative structure-activity relationship (QSAR) Tool (http://bci.clemson.edu/tools/glycan_array) used to analyze glycan array data.

The remainder of this thesis is organized as follows. Chapter 2 provides background and terminology information, including concepts to construct our new bioinformatics computing platform. Chapter 3 gives a detailed solution to design and deploy data-intensive computing platform for bioinformatics applications. Chapter 4 demonstrates the CMD project on HPC and virtualization environment. Chapter 5 shows a glycan array QSAR on our new bioinformatics platform. And finally Chapter 6 offers conclusions and future works.

Chapter 2

Background

Bioinformatics is an interdisciplinary field that combines aspects of Biology, Mathematics, and Computer Science [13], which are also theoretical foundations of our bioinformatics platform and related projects. In following sections, we introduce several terminologies and concepts in these areas to give a basic background for the further discussions in the remainder chapters.

2.1 Virtualization

Virtualization is a computing construct for running software (usually operating systems) concurrently and isolated from other programs on a single computer system [14]. Figure 2.1 shows the architecture of a virtualization platform. Typically, the architecture of OS virtualization includes a hypervisor, a software layer or subsystem that controls hardware and provides guest OSs with access to underlying hardware. The hypervisor allows multiple individual guest OSs to share the same physical system by offering virtualized hardware using various approaches (such as, full virtualization, para-virtualization and software virtualization) to mapping [15]. Guest OSs could be 32-bit or 64-bit Windows, Linux or Unix systems, which provides wide support for various applications and services. There are many different types of virtualization platforms, including Xen, KVM, VMware, VirtualBox, OpenVZ and Solaris containers.

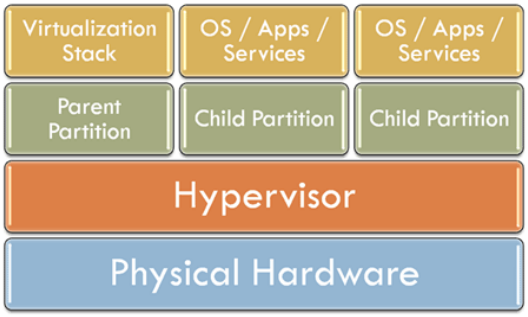


Figure 2.1 The architecture of virtualization platform

In our project, a Xen-based virtualization solution is used to support two product websites. The Xen framework is a very popular and common solution for the Linux platform. The Xen hypervisor contains three components, Domain 0, the Xen Hypervisor and Multiple Domain U [14]. These run directly on top of the physical machine, and act as the middle layer for the guest operating systems to access all hardware such as CPU, I/O, and disk. Domain 0 is the only domain that has privileges to access the Xen hypervisor. These privileges allow Domain 0 to manage and control all Domain Guests (DomUs), including starting, stopping, network requests, and so on.

2.2 Machine Learning

2.2.1 Classification and Prediction

Classification, also called supervised learning, is the computational approach of finding a model from training data and predicting the class of testing data. The model built from training data set has the ability to characterize and distinguish prediction data sets with a form of rules, decision tree or mathematical functions. Sample sets are usually observations or measurements labeled with features and class. Features selected for the

samples should relate to the classes, and the labels predicted by the classifier are usually categorical. For models generating continuous numerical values, the labels can be obtained by discretization. Some typical classification algorithms are decision tree, random forest, support vector machine and naïve Bayesian.

Various approaches can be used to validate the prediction result. The cross validation is one of the common and effective approaches. In this method, the sample data is randomly split into *n* sets. One set in the middle is used for testing and the other $n-1$ sets are used for training. Totally *n* iterations will be conducted, and the average of each result will be calculated to represent the performance.

2.2.2 Support Vector Machine

Support Vector Machines (SVM) [16] are a computational method for data classification by constructing a hyperplane (or set of hyperplanes) in a high or infinite dimensional space, which can be used for regression, classification, or other further tasks. The original problem is often stated in a finite dimensional space in which the sets to discriminate are not linearly separable. The principle of SVM is to map the original finite-dimensional space into a much higher-dimensional space to make the separation easier. A hyperplane created by the SVM is based on the largest distance to the nearest training data points of any class (functional margin). To avoid overfitting the problem, a buffer (soft margin) is used to provide certain error tolerances during the training stage. There are many implementations of SVM, such as LibSVM [17], SVMlight [18] and

TinySVM [19]. In our project, we use LibSVM tools with different kernel functions to improvement Simple Sequence Repeats (SSR) redundancy of the CMD website.

2.2.3 Evaluation Criteria

In order to measure the performance of classification result, multiple criteria were employed including sensitivity, specificity, precision, accuracy and F-measure. Four basic terminologies are used to definite these criteria which are true positive (TP), false positive (FP), true negative (TN) and false negative (FN) respectively. True positive is the correctly predicted positive data. True negative is the correctly predicted negative data. False positive is the predicted positive data that actually belong to the negative class. False negative is the predicted negative data that actually belong to the positive class. The definition of sensitivity, specificity, precision, accuracy and F-measure are shown here,

Sensitivity is the probability of correctly predicted positive data over the total number of positive data.

$$\text{Sensitity} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.1}$$

Specificity is the probability of correctly identified negative data over the total number of negative data.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FN}} \tag{2.2}$$

Precision is the probability of correctly predicted positive data over the total number of predicted positive data.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.3}$$

Accuracy is the probability of correctly predicted positive and negative data over the sum of positive and negative data.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2.4}$$

The F-measure or F-score is a measure of the accuracy of testing data by considering both the precision and recall of the test to compute the score. The best F-measure score is 1 and the worst F-measure score is 0.

$$\text{F} = \frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision+Recall})} \tag{2.5}$$

Additionally, ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) value serve to evaluate the discriminate power of models, and can be used to select the optimal models. ROC curve is used to measure the classifier in the cross validation study and the prediction performance. ROC curve plots the fraction of true positives against the false positive rate as the threshold of prediction varies. ROC evaluates the performance of classifiers based on the tradeoff between specificity and sensitivity. While ROC curve provides a visualization method to evaluate a classifier, AUC (area under the curve) score is widely used by providing a numeric value for the comparison of prediction performance. While an AUC value of 1 means perfect prediction, an area of 0.5 indicates random prediction. Most common models should have AUC values between 0.5 and 1.0. The higher the AUC value, the better the model. High AUC value means that lowering the threshold for the prediction only brings in limited false positive samples.

2.3 Partial Least Squares (PLS) Regression

The PLS regression has been widely used to model the relationship between responses and predictor variables [20]. For example, responses are the properties of chemical samples and predicator variables are the composition of chemicals. In our study, the response is the binding intensity of glycan chains to glycan-binding proteins and the predictor variables are the sub-trees extracted from glycan chains. Unlike general multiple linear regression, the PLS regression can handle strong collinear data and the data in which number of predictors is larger than the number of observations. The PLS build the relationship between response and predictors through a few latent variables constructed from predictors. The number of latent variables is much smaller than that of the original predictors. Let vector y (n×1) denote the single response; matrix X (n×p) denote the n observations of p predictors and matrix T (n×h) denote n values of the h latent variables. The latent variables are linear combinations of the original predictors:

$$T_{ij} = \sum_{k} W_{kj} X_{ik} \tag{2.6}$$

where matrix W (p×h) is the weights. Then, the response and observations of predictors can be expressed using T as follows [20]:

$$X_{ik} = \sum_{j} T_{ij} P_{jk} + E_{ik} \tag{2.7}$$

$$y_{m} = \sum_{j} C_{mj} T_{ij} + f_{m} \tag{2.8}$$

where matrix P (h×p) is the is called loadings (the regression coefficients of latent variables T for observations) and matrix C (h×1) is the regression coefficients of T for responses. The matrix E (n×p) and vector f (n×1) are the random errors of X and y. The

PLS regression decomposes the X and y simultaneously to find a set of latent variables that explain the covariance between X and y as much as possible [20].

The PLS regression was performed using the plsregress function in Matlab. The plsregress function takes three parameters: X, y and the number of components. It is important to determine the number of components in PLS regression. We employed the following procedure to select the number of components. We first ran the PLS regression using a large number of components, for example, 50. The plsregress returned the percentage of variance explained in response for each PLS component. Then, we counted the number of components that contribute to variance explained beyond a threshold. This number was our new number of components. In our study, we set the threshold to be 0.5% of variance explained. We then ran PLS regression again using the new number of components.

The R^2 of PLS regression is calculated using the formula: $R^2 = SS_{err} / SS_{total}$. The SS_{err} is the sum of squares of fit errors: $SS_{err} = \sum_i f'_i$, where f' (n×1) is the regression errors. And the SS_{total} is the total sum of squares: $SS_{total} = \sum_i (y_i - \bar{y})^2$, where $\bar{y}$ is the mean of y.

2.4 GMOD

Generic Model Organism Database (GMOD) project, is a collection of open source software tools to create and manage genome-scale biological databases [21]. There are more than 37 components and 14 functionality areas in GMOD project. These components provide the functionality that is needed by all organism databases like Community Annotation, Comparative Genome Visualization, Database schema, Database

tools, Gene Expression Visualization, Genome Annotation, Genome Visualization & Editing, Ontology Visualization, Literature Tools, Workflow Management, Molecular Pathway Visualization, Middleware, Tool Integration and Sequence Alignment.

In modern biology research area, bioinformatics applications and databases are begin developed at a steady rate. However, many of these tools are seldom used since the user may not have the resources or skills to install the tool and integrate them. There is need for a standardized solution to integrate those tools and databases together. GMOD provides such a platform for developers, scientists and laboratories to construct their own bioinformatics software.

2.5 BLAST and FASTA

BLAST (Basic Local Alignment Search Tool) program [22] and FASTA program [23] are both tools for sequence similarity search which can be used to compare a query to a DNA/protein database by a stand-alone tool or a web interface. The difference between BLAST and FASTA is that they use different algorithm for comparison. FASTA is better for less similar sequences. BLAST may be faster than FASTA without significant loss of ability to find the similar sequences in the DNA/protein database. BLAST is one of the most widely used bioinformatics tools. There are several variants of BLAST programs that can compare between protein or nucleotide queries with protein or nucleotide databases.

2.6 BioPerl

BioPerl is an open-source international project (since 1995) [5] to facilitate sequence alignments, genetic sequence manipulation and genomic analysis. It allows bioinformatieists, genetic researchers and computer scientists to collaboratively focus on providing a set of well-documented Perl modules [24]. It provides a set of foundational libraries that allow the building of complex bioinformatics tools for use in production quality software [4] and the construction of complex solutions to bioinformatics problems. BioPerl (version 1.6.9) gives support to read/write of multiple sequence file formats, sequence retrieval from web databases, sequence manipulation and alignment and sequence annotations.

2.7 Drupal

Drupal is one of most widely used open source web Content Management Systems (CMS), which are used to create integrated web sites. Drupal web sites can include a blog, a portal web site for the organization, an e-commerce site, a social networking site and other componets [25]. The framework of the Drupal system is highly modular, extensible, and standards-compliant. The official version of Drupal only contains the basic core functionality, however, additional functionality can be added using built-in or third-party modules. There are more than 9,000 modules available to extend and customize Drupal functionality. To apply these modules does not require any modifications to the code in the core.

CMS use in bioinformatics is growing [26]. Drupal is used for many applications, including in on-line analysis tools, intranet tools, collaboration tools, biology databases,

conference websites and lab websites. For example, Drupal is selected as the core development framework in GMOD Tripal project [27]. The Tripal modules allow the Drupal CMS to interact with Chado [28] data, as well as provide data loaders, display of Chado data and administrative interfaces for data management. Our glycan array QSAR tool also was developed as a module of Drupal platform by customizing Webform module and Theme module. Detailed discussion will be given at the next chapter.

Chapter 3

Bioinformatics Platform Design, Implementation and Deployment

In this chapter, we will discuss the design, implementation and deployment of the system platform regarding data-intensive bioinformatics applications. The primary design goal is to create a flexible, configurable and high-performance framework with a user-friendly web interface. Based on the new design, biologists with minimal computer science background can easily analyze massive data sets by using public computing infrastructure. Figure 3.1 shows the architecture of the data-intensive computing platform for bioinformatics applications using virtualization technologies and HPC infrastructures. Detailed description of the multi-tier architecture for the bioinformatics platform is presented in the following sections of the thesis.

3.1 Multi-tier Architecture

Our design applies the software engineering concept of multi-tier architecture including presentation tier, logic tier and data/computing tier. It can seamlessly integrate the web user interface, scientific workflow and computing infrastructure (Figure 3.1).

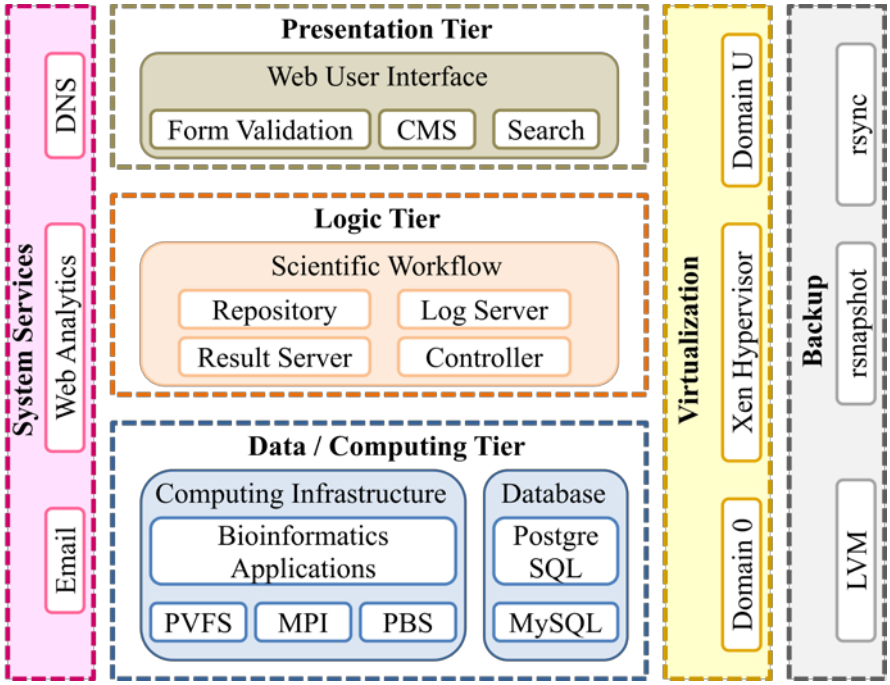


Figure 3.1 The architecture of data-intensive computing platform for bioinformatics applications.

Presentation tier

The presentation tier is the top layer of the platform. The presentation tier provides user-friendly web interface related to various bioinformatics applications, such as tools for sequence alignment in similarity comparisons, protein sequence or nucleic acid assembly, genetic map views, SSR information retrieval, or glycan array data analysis. All computational parameters and input data sets are submitted from this tier, and then initial information is passed to the next tier. The presentation tier includes three services: a form validation service, the CMS service and a search service. The interface of this tier is implemented using HTML, Perl, PHP, JavaScript languages and the Drupal CMS module.

Logic tier

The logic tier is a middle layer between the presentation tier and the data/computing tier. It acts as a workflow control center for the whole system. The logic tier accepts task requests from the front web user interface and generates the Portable Batch System (PBS) scripts for the HPC infrastructure based on the resource requirements of the bioinformatics applications. In necessary, the logic tier transfers input data sets and PBS scripts to the HPC infrastructure. It submits tasks to the task scheduler of the HPC infrastructure, monitors status of the tasks, collects and stores task results on the web server, parses and generates user-friendly results (e.g. Microsoft Excel format), and sends an email to users with a summary page linked to the task directory on the web server.

The logic tier is composed of four services: the controller, the repository, the result server and the log server. The controller manages the life cycle of each task, transfers, and submits all tasks to the remote HPC infrastructure. The repository is a file archive service. It maintains a unique working directory for each task. All files including input data sets, parameter files, PBS scripts, output result files, parsed results and log files, are hosted in their corresponding directories. The result server receives the properly processed results and sends the confirmation email to the users. If any unexpected exception happens, the result server sends an error message to users and the system administrator. Finally, the log server records the execution details for each task, providing a real-time feedback to monitor the service quality of the current system.

Data/Computing tier

This tier is the bottom layer of the platform. It consists of storage and computing resources including five components: the database system, the PVFS file system, the MPI (Message Passing Interface) library, the PBS job scheduler and bioinformatics applications. The database system stores the relational data sets presented on the bioinformatics website. The PVFS file system can support the management for petabyte-scale massive files [9], where the large amounts of protein databases and computational intensive applications are hosted. The MPI library provides the fundamental parallel mechanism for bioinformatics applications to achieve the high-performance and scalability. The PBS job scheduler controls batch jobs and distributed computing resources. Bioinformatics applications are installed on the shared storage system which allows all computing nodes to access the same version of program.

3.2 Presentation Tier

3.2.1 CMD Web Interface

Home page

The user interface of the CMD website consists of four functional areas including the menu area, the navigation area, the search area and the content area (Figure 3.2). The menu area includes four main sections: General Info, View, Search and Resources. Each has a drop down menu at the top of the home page, with further options seen along the grey bar. The user can quickly access needed sections by moving pointer over the green title of the link from the drop down menu, and they also can get the section name from

navigation area. The search bar at top of the every page is a convenient way for a user to retrieve the related information from whole website.

Figure 3.2 Home page of the CMD website. A. The menu area. B. The navigation area. C. The search area. D. The content area.

The primer redundancy page

The primer redundancy project page in CMD displays the results obtained from the analysis. The summary page explains the type of analysis performed, number of primers used, total number of redundant primers, threshold value for the analysis and the type of redundancy. The redundant primers page provides a list of redundant primers found in the CMD. The primer redundancy web interface can be accessed at http://www.cottonmarker.org/primer_redundancy/. Figure 3.3 shows a query result page of SSR 'BNL3500' where the user can explore all redundancy information related to the current SSR.

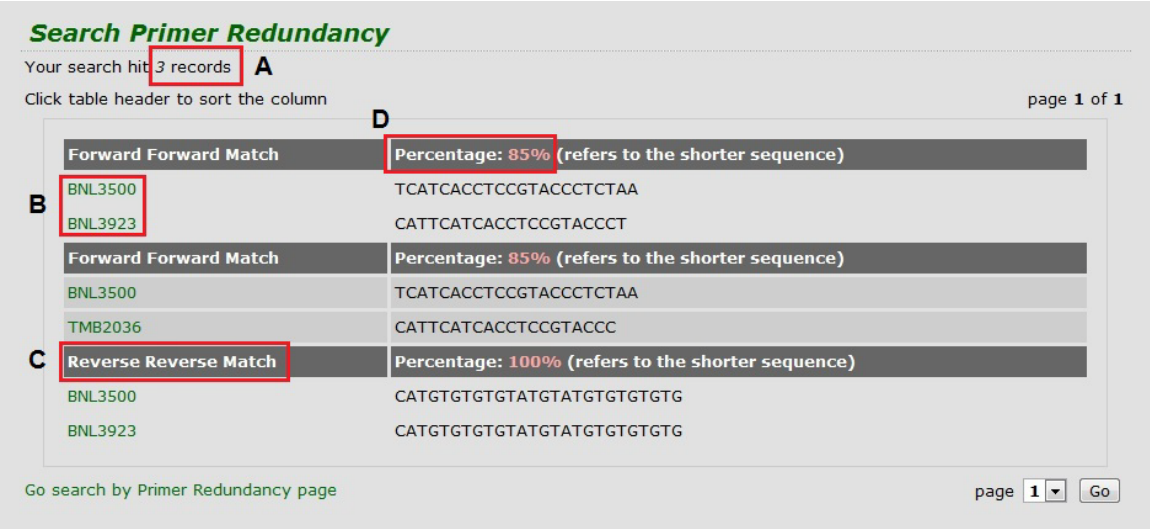


Figure 3.3. The web interface for primer redundancy page. A. The number of matched records. B. The name of the matched primer pair. C. Match type. D. Similarity of the mached pair

Traits page

The traits page visually displays and compares linkage groups/chromosomes of each cotton genetic cross available in CMD, including Quantitative Trait Locus (QTLs)

associated with the cross, as well as their exact positions on the respective chromosomes. Overall, all CMD features are intertwined, offering simple and easy access to published data related to cotton molecular breeding. For example, from the View Traits page (Figure 3.4A), one can gain access to published information about the trait-associated QTLs and nearest mapped SSR markers (Figure 3.4B), as well as detailed information about each QTL (Figure 3.4C) and associated marker information (Figure 3.4D).

Genetic map (CMap) page

CMap is part of GMOD [29] which allows the user to browse the map of interest and select other maps for comparisons. Users can view the number of correspondences among all selected maps in the CMap correspondence matrix. The feature search looks for a certain feature by name or species, accession ID and feature type. Currently, CMD contains data for 27 genetic maps. The anchored genetic markers can be viewed in several formats which includes an Excel spreadsheet, a database search interface, and a graphical interface for comparative SSR maps. Figure 3.6 shows the graphical interface of CMap in which the cotton genetic maps are displayed with anchored SSR markers, and the SSRs location is also compared between different crosses of cotton.

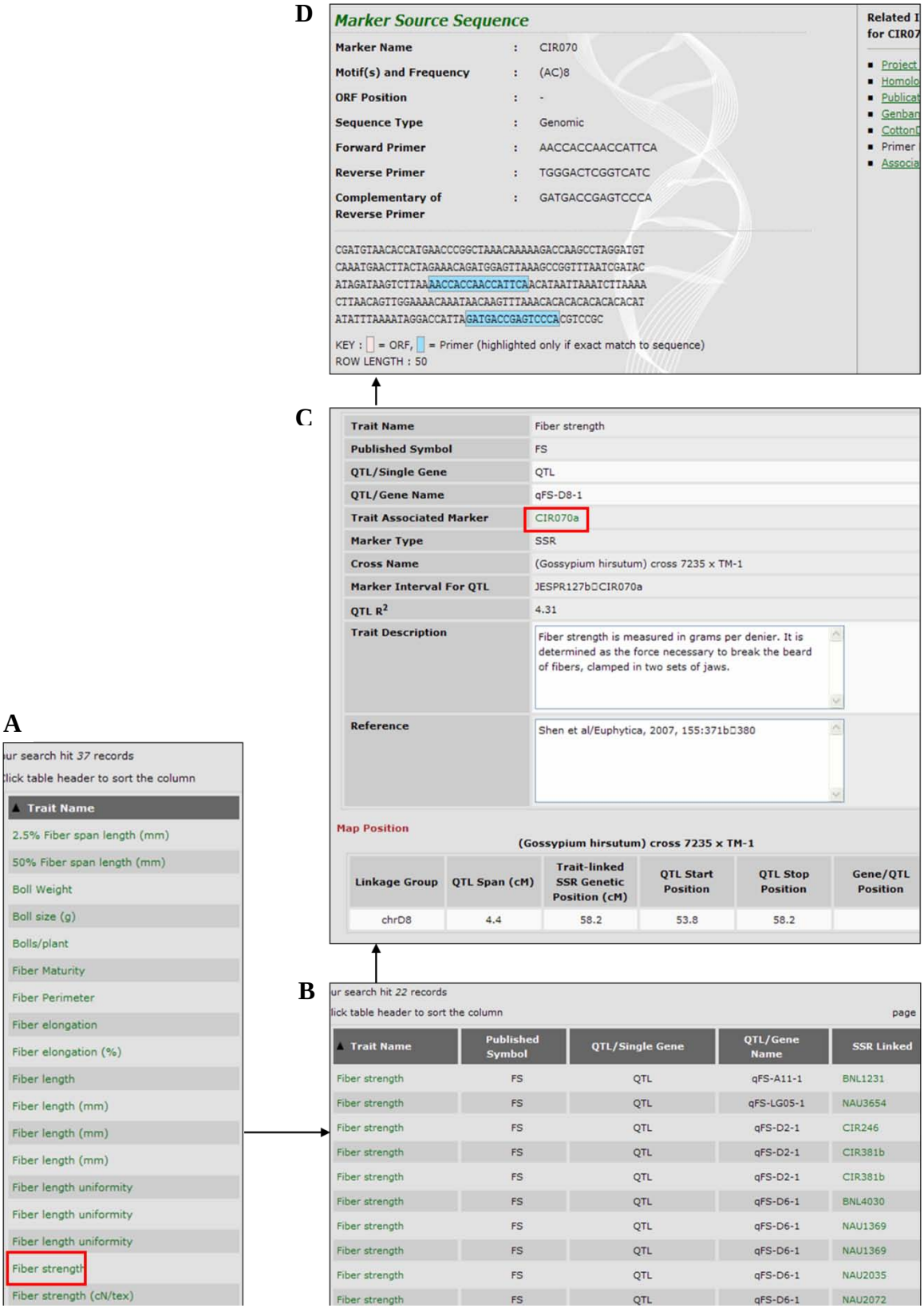

D
Marker Source Sequence
Marker Name : CIR070
Motif(s) and Frequency : (AC)8
ORF Position : -
Sequence Type : Genomic
Forward Primer : AACCACCAACCATTCA
Reverse Primer : TGGGACTCGGTCATC
Complementary of Reverse Primer : GATGACCGAGTCCCA
CGATGTAACACCATGAACCCGGCTAAACAAAAAGACCAAGCCTAGGATGT
CAAATGAACTTACTAGAAACAGATGGAGTTAAAGCCGGTTTAATCGATAC
ATAGATAAGTCTTAAAACCACCAACCATTCAACATAATTAAATCTTAAAA
CTTAACAGTTGGAAAACAAATAACAAGTTTAAACACACACACACACACAT
ATATTTAAAATAGGACCATTAGATGACCGAGTCCCACGTCCGC
KEY : = ORF, = Primer (highlighted only if exact match to sequence)
ROW LENGTH : 50
Related I
for CIR07
Project
Homolo
Publicat
Genban
Cotton
Primer
Associa
C
Trait Name Fiber strength
Published Symbol FS
QTL/Single Gene QTL
QTL/Gene Name qFS-D8-1
Trait Associated Marker CIR070a
Marker Type SSR
Cross Name (Gossypium hirsutum) cross 7235 x TM-1
Marker Interval For QTL JESPR127b CIR070a
QTL R2 4.31
Trait Description Fiber strength is measured in grams per denier. It is determined as the force necessary to break the beard of fibers, clamped in two sets of jaws.
Reference Shen et al/Euphytica, 2007, 155:371b 380
Map Position
(Gossypium hirsutum) cross 7235 x TM-1
Linkage Group | QTL Span (cM) | Trait-linked SSR Genetic Position (cM) | QTL Start Position | QTL Stop Position | Gene/QTL Position
chrD8 | 4.4 | 58.2 | 53.8 | 58.2
A
ur search hit 37 records
lick table header to sort the column
Trait Name
2.5% Fiber span length (mm)
50% Fiber span length (mm)
Boll Weight
Boll size (g)
Bolls/plant
Fiber Maturity
Fiber Perimeter
Fiber elongation
Fiber elongation (%)
Fiber length
Fiber length (mm)
Fiber length (mm)
Fiber length (mm)
Fiber length uniformity
Fiber length uniformity
Fiber length uniformity
Fiber strength
Fiber strength (cN/tex)
B
ur search hit 22 records
lick table header to sort the column
page
Trait Name | Published Symbol | QTL/Single Gene | QTL/Gene Name | SSR Linked
Fiber strength | FS | QTL | qFS-A11-1 | BNL1231
Fiber strength | FS | QTL | qFS-LG05-1 | NAU3654
Fiber strength | FS | QTL | qFS-D2-1 | CIR246
Fiber strength | FS | QTL | qFS-D2-1 | CIR381b
Fiber strength | FS | QTL | qFS-D2-1 | CIR381b
Fiber strength | FS | QTL | qFS-D6-1 | BNL4030
Fiber strength | FS | QTL | qFS-D6-1 | NAU1369
Fiber strength | FS | QTL | qFS-D6-1 | NAU1369
Fiber strength | FS | QTL | qFS-D6-1 | NAU2035
Fiber strength | FS | QTL | qFS-D6-1 | NAU2072

Figure 3.4 CMD Trait View Page. A. Full traits list page. B. The page for the same trait with different properties (QTL/Gene Name, SSR Linked). C. The representative trait information page. D. The page for associated marker.

Figure 3.5 CMD CMap Viewer. A). The index page of CMap Viewer with 27 major genetic maps of cotton. B). An example (chromosome 24) displayed in CMap Viewer with the associated mapped QTLs (light-green bars represent QTLs with their names on the left).

3.2.2 BCI Glycan Array QSAR Tool Interface

To facilitate the utilization of our QSAR method by biologists to analyze their own glycan array data, we developed a web tool and hosted it at http://bci.clemson.edu/tools/glycan_array. The web interface as shown in Figure 3.6. The

users first need to choose three parameters: the array version, sub-tree features and z-score for selecting significant sub-trees. They then need to paste a one-column binding intensities of glycan array. After clicking the “Submit” button, the parameters and data are transferred to the server. A MATLAB program on the server side will perform the PLS regression and generate significant sub-trees. The server will then generate a results page and send back to client. As shown in Figure 3.7, the results page contains a summary section of input parameters and R^2 value; a table of the significant sub-trees, their regression coefficients and glycan chains containing each feature; a figure that plots the percentage of variance explained against number of PLS components and a figure that plots the observed intensities against fitted intensities. The user will be able to download results and figures from the results page.

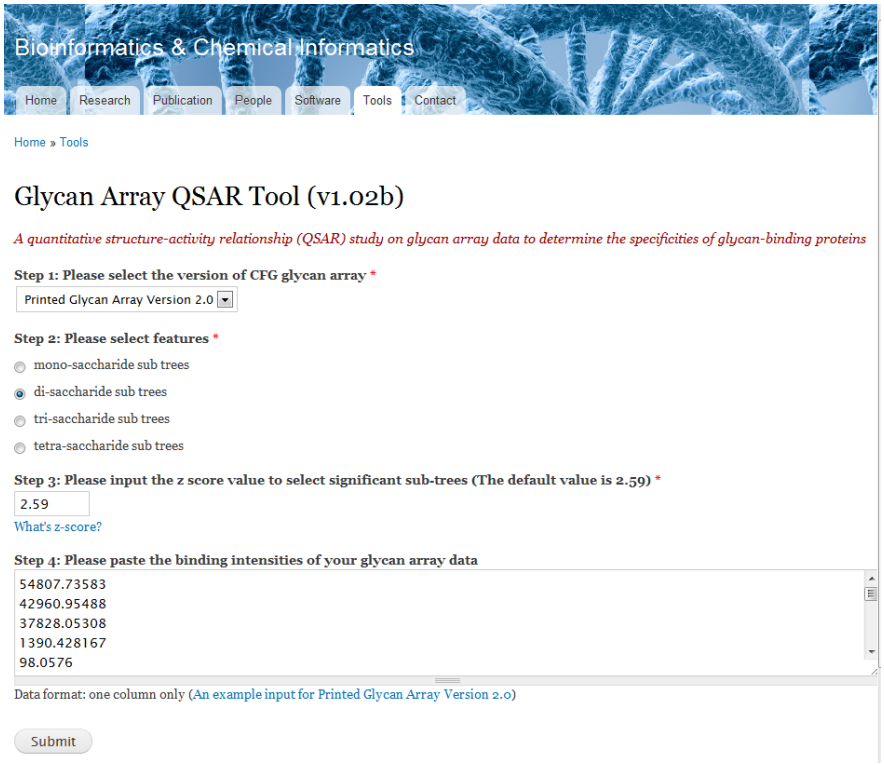


Figure 3.6 The web interface of Glycan Array QSAR Tool.

Glycan Array PLS Regression Results

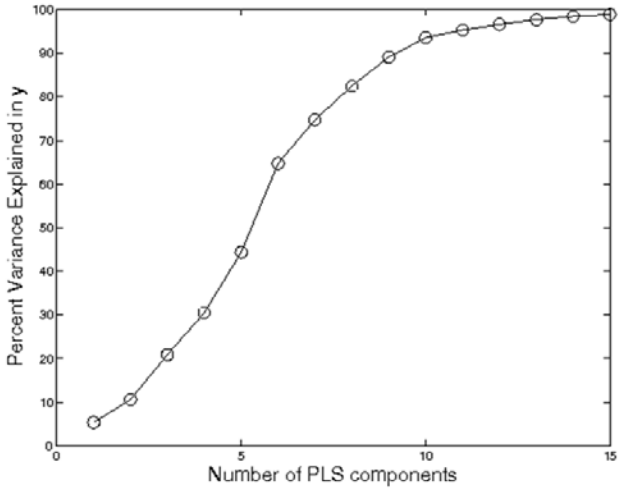
Summary

Printed Glycan Array Version: 4.1
Feature: mono-saccharide sub trees
Threshold for significant feature: 2.59
Rsquared: 0.988163

Significant Features	Regression Coefficients	Glycan Chains Contains the feature
3(6OSO3)Galb1	8655.456148	45
3(6-O-Su)Galb1	7816.967003	228

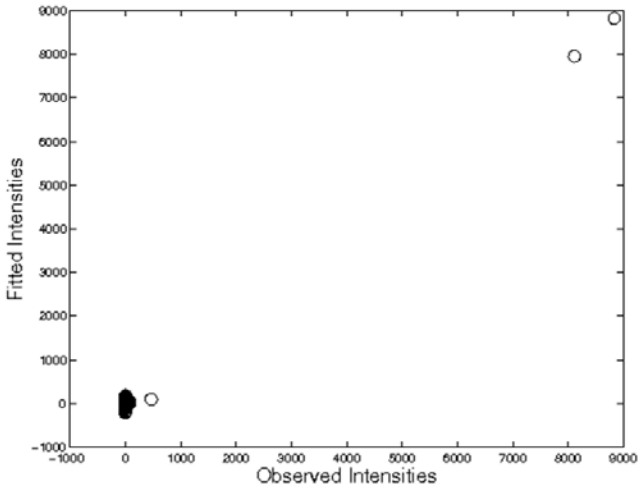
Download data file: Excel/csv format

Plot of percentage of variance explained against number of PLS components



Download

Plot of observed intensities against fitted intensities



Download

Figure 3.7 An example result of Glycan Array QSAR Tool showed the significant mono-saccharide sub-trees binding specifically to Siglec-8.

3.3 Logic Tier

3.3.1 Data Flow on Bioinformatics Platform

The web interface is the gateway between the user and bioinformatics computing resources. The initial information is sent from web interface and finally returned to the front web interface again. Figure 3.8 illustrates the lifecycle of data flow on the CMD system. The computing jobs are requested by the user from the server tools page in the CMD website. The controller in the logical tier schedules the current computing job to local computing server or remote HPC infrastructure based on the computing resources which the user requests. To execute large-scale computing jobs, the user request (including input files, parameters and an execution script) is transferred to the remote PBS scheduler that directs the job to the HPC infrastructure. When the computing job is finished, the controller will either send error message to the user and system administrator or pass the original output (raw data) to the result server.

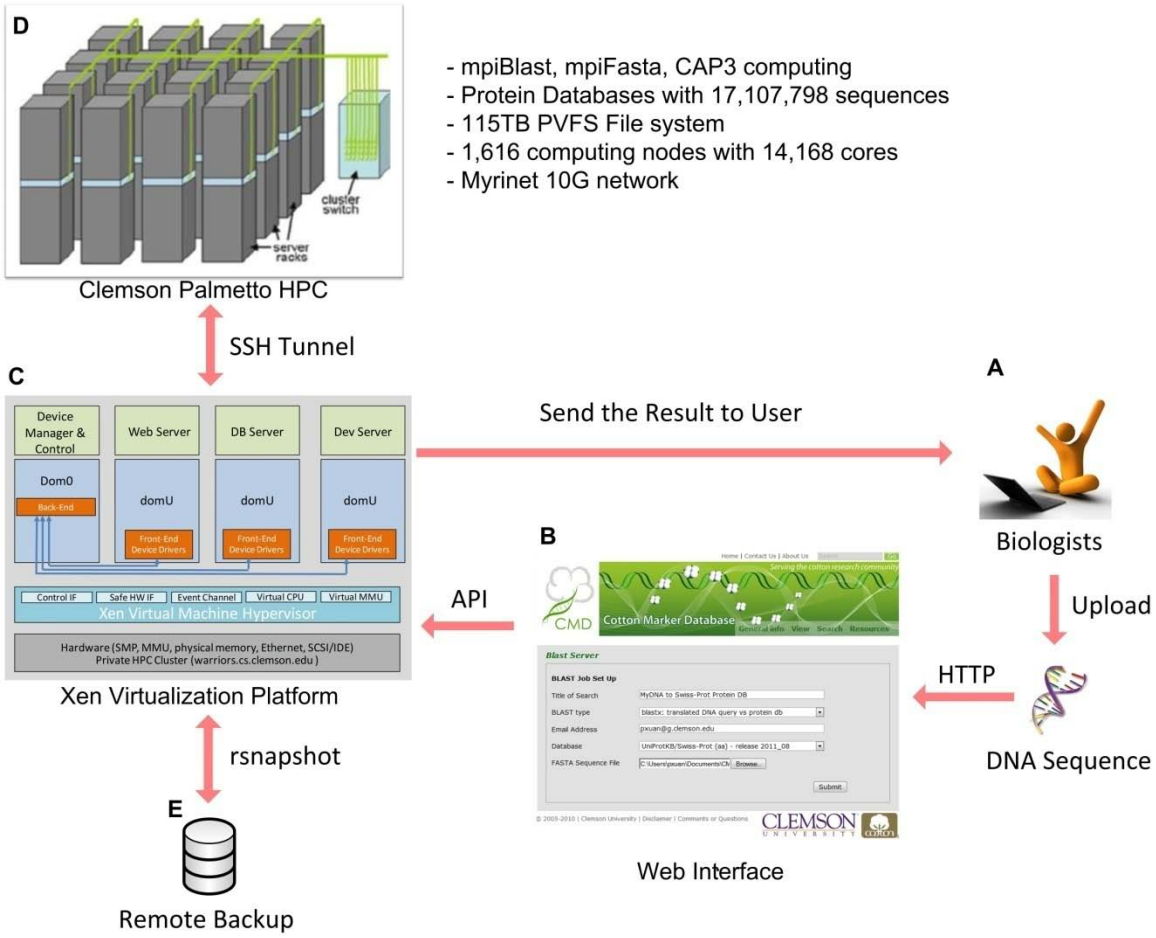


Figure 3.8 The data flow on Cotton Marker Database. (A). Biologists retrieve the genetic information or submit job from the thin client side. (B). CMD user-friendly web interface. (C). Xen virtualization solution for all servers. (D). All data-intensive computing jobs are transferred to Palmetto HPC. (E). Remote snapshot can quickly mirror all virtual machines to another location.

3.3.2 CMD Web Tools

The tools page in the CMD website provides access to a CAP3 Assembly server, an SSR server, and BLAST and FASTA sequence similarity search tools. These jobs are both time- and resource-consuming, and very difficult to process on a single machine. There are three protein databases with a total size of 16GB that need to be compared

during each computational job. The average computational time could take more than two weeks depending on the size of query file user submitted. In order to handle the larger size of sequence and also reduce job running time, we redesigned the job execution pipeline, and transfer most computational tasks to the Palmetto HPC platform. The Figure 3.9 illustrates the detailed workflow for job execution in the CMD platform.

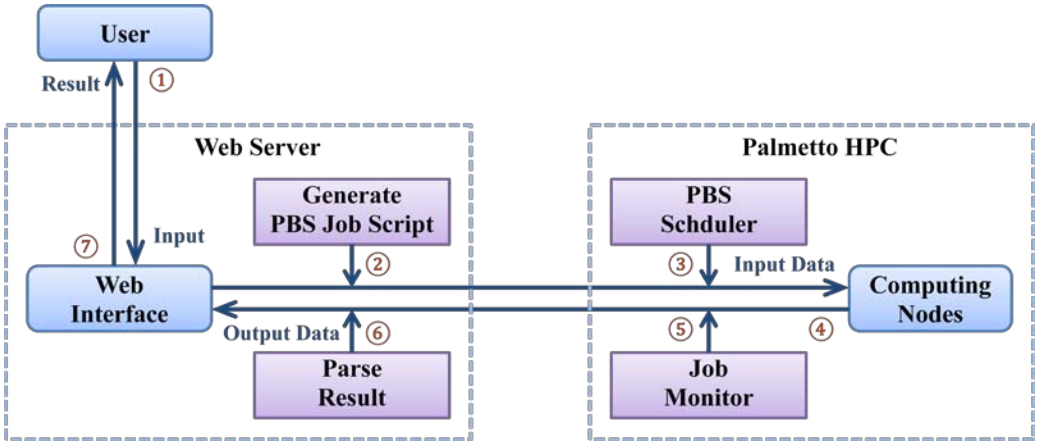


Figure 3.9 CMD web tools execution workflow between the web server and Palmetto HPC.

CMD web tools execution workflow including the following steps:

1. The user uploads the input file and parameters to analysis tools on CMD website.
2. Web server generates the PBS script based on the computational job requested by the user and then transfers all initial files (input file, PBS file, and parameter file) to Palmetto HPC using scp.
3. All jobs are submitted to PBS scheduler of HPC.
4. Jobs are distributed to computing nodes.

5. The job monitor keeps tracking the status of each job and sends the result back to CMD web server after each job is done.
6. The result server parses the output file to generate the result file with Excel format.
7. CMD web server sends an email with a link to the result page.

SSR server

SSR analysis tools are implemented using a Perl script SSRIT [30] and the FLIP [31] program of the Organelle Genome Megasequencing Project [32]. The information of Open Reading Frames (ORFs) is extracted by the FLIP program, and potential primers are identified using Primer3 [33]. Users can submit a SSR job using the web-based interface of the SSR server by uploading a batch of sequences with FASTA format and related parameters. The result is presented by a job summary page and users also receive an email which includes a URL link to the summary page of this job, which includes:

a) a report of the SSR analysis
b) a library file of sequences user submitted
c) a library file of the SSR containing sequences
d) an Excel file including the SSR-containing clones and its individual properties
e) detailed properties page including sequence name, repeat(s) motif and number, length of the SSR-containing sequence, ORF start/stop position, SSR start/stop position, SSR location relative to the ORF, primer pairs and GC content of the sequence.

CAP3 server

We have deployed the Contig Assembly Program called CAP3 [34] on the bioinformatics platform to allow users to assemble ESTs. Users can submit quality files of sequences with the percentage identity in the overlap region using the web interface. As more ESTs are available in the public database, the unigene of cotton can be continually refined via our CAP3 server, and more SSRs could be mined via the SSR server.

Sequence similarity servers

CMD FASTA and BLAST servers allow users to perform homology batch queries between their sequences and the cotton SSR sequences or protein databases by user-friendly web interface in CMD. In the result summary page, users can retrieve an original output file and an Excel file including any known function, the best match, match length, alignment length, match organism, percent identity, expectation value, and start and stop alignment positions. Our sequence similarity servers specifically designed for processing the larger-scale jobs in cotton research areas, which can help researchers compare new developed cotton sequences and reduce potential redundancy when developing new markers.

Web query services

The database query service is implemented using Perl language, JavaScript language, CPAN and BioPerl modules. Perl is a very popular and also regular programming language in the biology area because it is so well-suited to several

bioinformatics tasks. It is easier to use, more efficient and powerful than traditional programming languages [35]. All query pages are dynamically generated by Apache Perl CGI [36] module, and the result is extracted using SQL language from database. The concept of object oriented [37] design is applied to the development of CMD website.

3.3.3 BCI Glycan Array QSAR Tool

The Glycan Array QSAR Tool provides a novel quantitative structure-activity relationship (QSAR) method to analyze the glycan array data. It first decomposes glycan chains into mono-, di-, tri-, or tetra-saccharide sub-trees. The bond information is also incorporated into the sub-trees to help distinguish the glycan chains structurally. Then, the tool performs PLS regression on glycan array data using the sub-trees as features. Based on the regression coefficients of PLS, the tool will report the sub-trees that determine the binding specificities of glycan-binding proteins. Figure 3.10 illustrates the detailed job execution workflow for this tool.

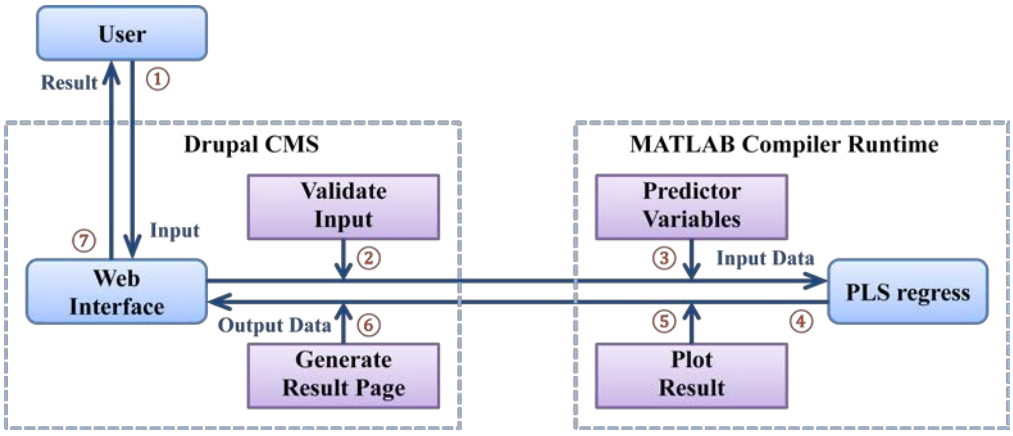


Figure 3.10 The BCI Glycan QSAR tool execution workflow between web server and Palmetto HPC.

BCI Glycan QSAQ tool execution workflow includes the following steps:

1. The user inputs parameters and glycan array data from the web interface.
2. The form program validates the correctness of input.
3. The corresponding predictor variables are passed to the regression program.
4. Perform the MATLAB PLS regression program and generate significant sub-trees.
5. Plot figures.
6. Generate the result page.
7. Send result page to user.

3.4 Data / Computing Tier

3.4.1 Database System

There are two kinds of database management systems to serve the CMD website. The CMD main website is a relational database implemented by MySQL open source database version 5.0.77. Currently, the database of the CMD website contains 23 tables that host all the data set for the SSR projects, which include information on sequences, SSR-containing clones, primers flanking the SSRs, repeat motif, trait, project collaborators, genetic markers and maps, open reading frame position, standardized panel varieties, publications, data homology and primer redundancy. Figure 3.11 shows the database schema of the CMD website.

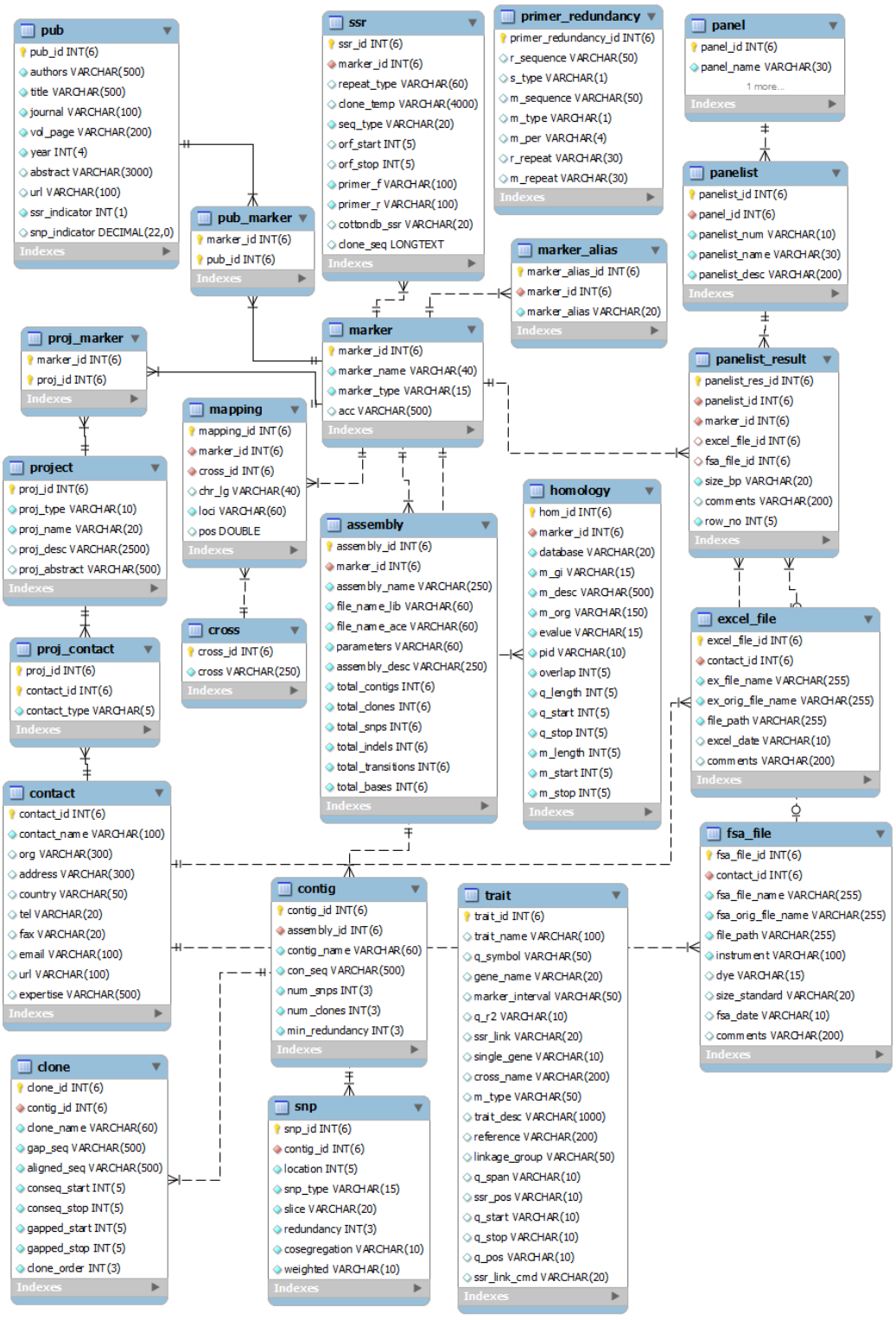


Figure 3.11 The database schema of CMD website

3.4.2 Directory Structure on Palmetto HPC

Both data-intensive and computational-intensive bioinformatics jobs are transferred to Palmetto HPC. A hierarchical directory structure (Figure 3.12) is used to organize files (protein databases, bioinformatics applications, computing jobs, scoring matrices and utilities) on Palmetto HPC.

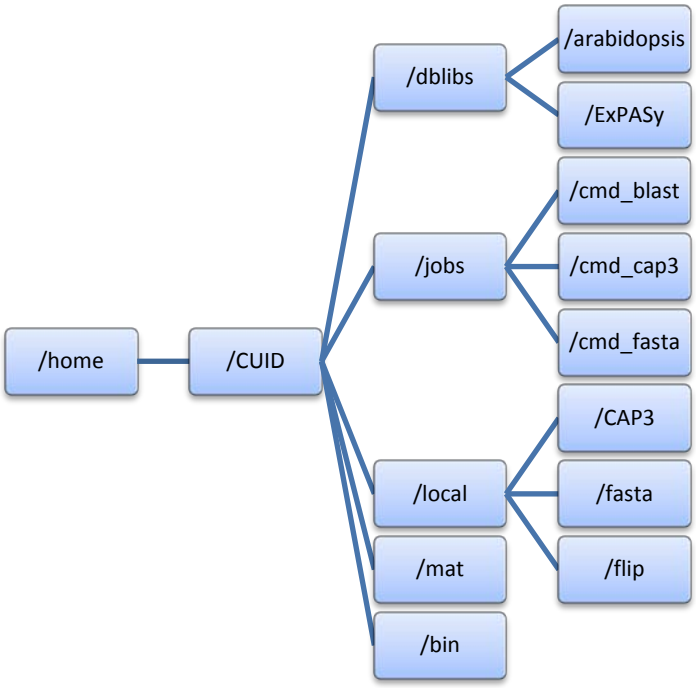


Figure 3.12 The directory structure in Palmetto HPC

dblibs

All databases (totally 17,107,798 protein and DNA sequences) used by BLAST, FASAT and CAP3 server are stored in this directory. When a new version of update is released, our update script can automatically download and preprocess the new database file.

jobs

We deploy several popular bioinformatics applications on Clemson HPC platform. Each computing job is assigned a unique job ID as the directory name of a working space, where input, output and log files is archived.

local

Packages and programs of bioinformatics applications are installed in this directory. These programs are invoked by PBS job script following with detailed parameters and requested computing resources.

bin

This directory hosts customized utilities and scripts. For example, we develop ‘submission.sh’ utility to monitor the status of computing jobs by wrapping PBS ‘qsub’ tool. This script can hold job submit terminal and does ‘polling search’. When job finish, the ‘submission.sh’ script can release job submit terminal. By this approach, the workflow control module of CMD system can get a signal of job execution status (waiting, running, success or failure) from the remote computing infrastructure.

mat

All protein similarity matrixes needed by BLAST and FASTA programs stores in this directory.

3.5 System Services

3.5.1 Email System

The mail system is a major channel to build the communication between the user and websites. Technically, we use Google Groups in Google Apps as the mailing list

system, and invoke Perl CPAN module (Net::SMTP::TLS) to send the email via the authentication SMTP protocol from the web server. Figure 3.13 shows the management interface of mail list system. Below are two mail list groups for CMD project:

- cmd_list@cottonmarker.org (52 users)
- cmd_advisory@cottonmarker.org (22 users)

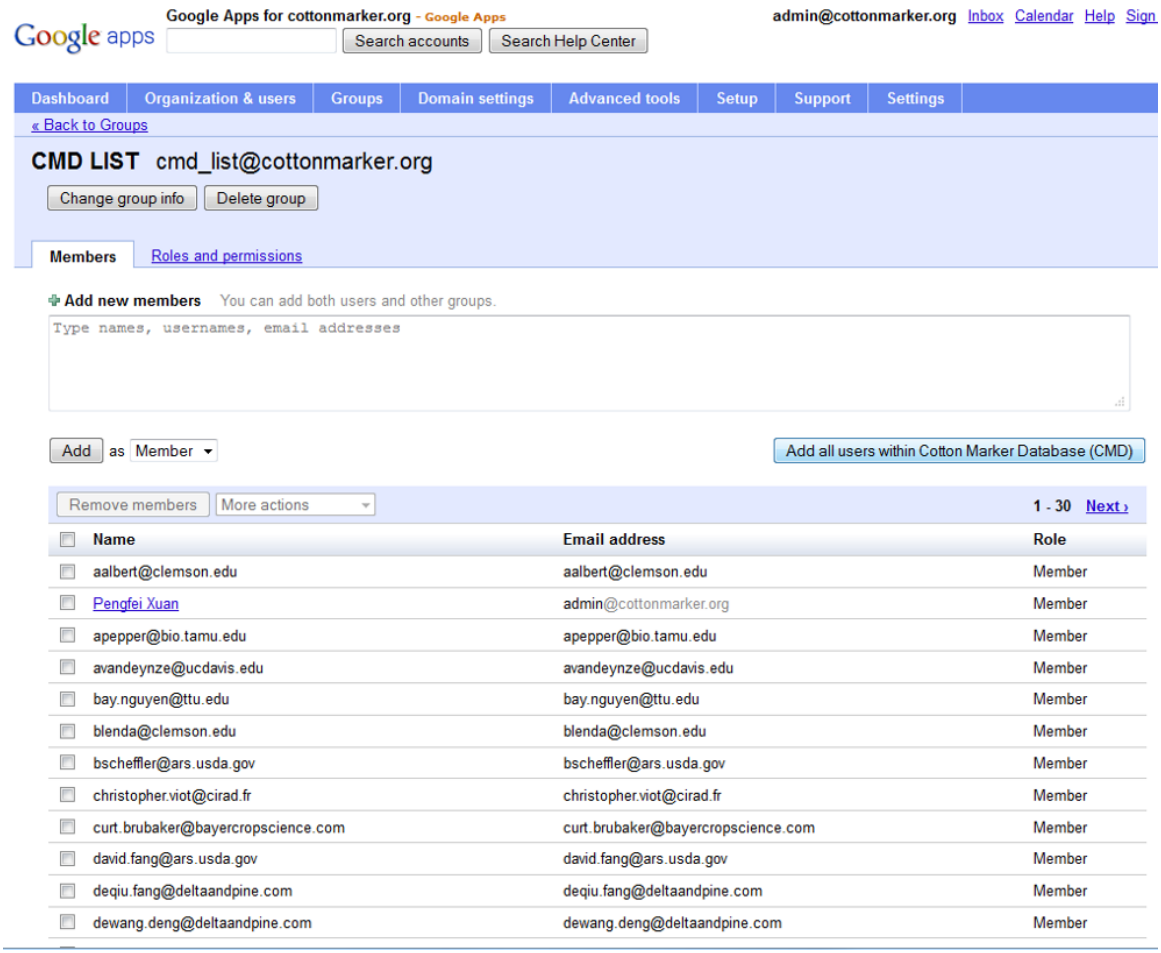


Figure 3.13 The management interface of CMD mail list system

3.5.2 DNS Service

DNS service of whole system is hosted at Clemson CCIT department. The domain name (www.cottonmarker.org) is pointed to the IP address of CMD VM web server (130.127.48.247), and the mail exchange (MX) records is pointed to Google Apps mail servers. Table 3.1 gives the configuration information for MX record.

Table 3.1 MX records of CMD mail servers

MX Server address	Priority
ASPMX.L.GOOGLE.COM.	10
ALT1.ASPMX.L.GOOGLE.COM.	20
ALT2.ASPMX.L.GOOGLE.COM.	20
ASPMX2.GOOGLEMAIL.COM.	30
ASPMX3.GOOGLEMAIL.COM.	30
ASPMX4.GOOGLEMAIL.COM.	30
ASPMX5.GOOGLEMAIL.COM.	30

3.5.3 Web Analytics

Web analytics is very important feature to the website, as it provides the collection, measurement, analysis and reporting of website access data that allow us to understand and optimize web usage [38]. In addition to measuring the website traffic, the web analytics can also be used as a tool to track the user behavior of accessing website. The Google Analytics [39] is a major engine used to track the access history for our websites. Figure 3.14 shows the interface of Google Analytics, which provides information about the number of page views and the number of visitors to a website; it generates traffic and popularity trends which are useful for CMD user distribution research; it also shows us how visitors found CMD website and how they interact with it. All of this information is presented in intuitive, thorough, visual reports. Web access log

files are archived at servers in Google instead of our server.

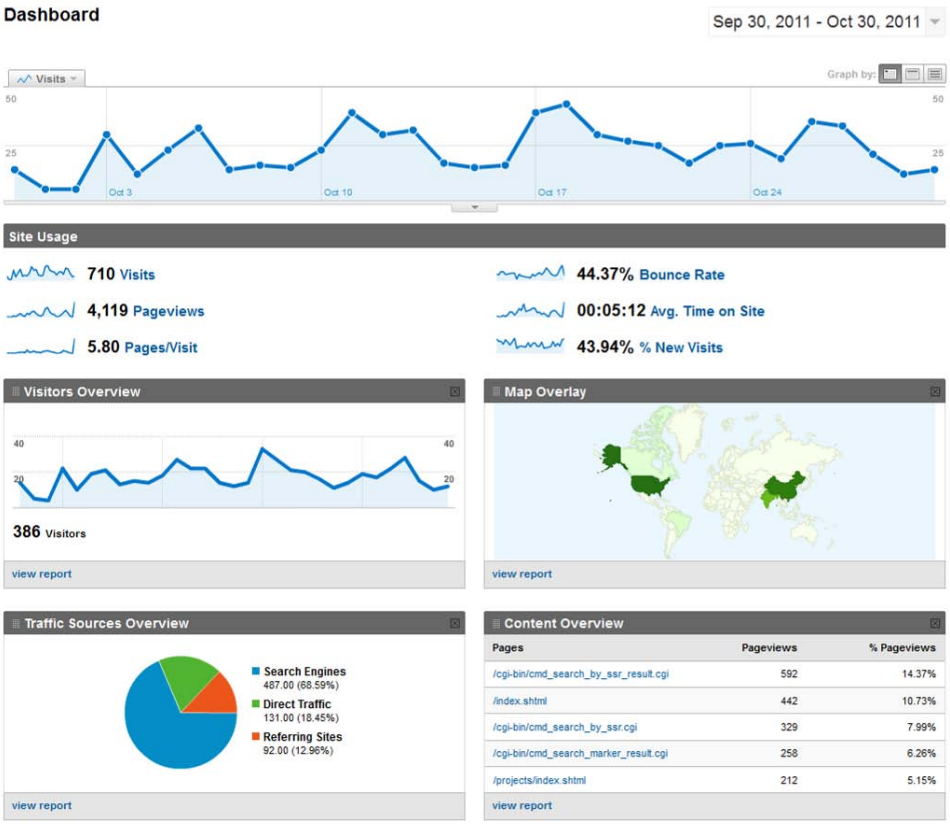


Figure 3.14 The interface of Google Analytics. It provides site usage, visitors, geography location, traffics sources and records of each page

3.6 Server Virtualization

The presentation and logic tiers of bioinformatics platform are constructed using a CentOS 5.4 Xen-based virtualization solution. We create three VMs (CMD web server, BCI web server and database server) at the top of physical server. This design allows use of separate different services at the OS level (Figure 3.8C). The physical server use Logical Volume Manager (LVM) to maintain the file system. The guest VM use Logical Volume (LV) of physical server as the local disk. LVM allows the dynamical extension of partition space without destroying the whole file system. The guest VMs connect to

the public network though the network bridge between domain 0 and domain U. All packets sent or received from VMs will pass through the PREROUTING, FORWARD and POSTROUTING iptables chains of domain 0, in which we can apply the global security policy to monitor and filter the network traffic for all VMs.

The server virtualization provides a lot of benefits for our bioinformatics platform. It increases the flexibility of deployment, reduces the complexity of platform design, improves the reliability of the platform and enhances the security of the whole system. For example, we can easily resolve the dependency conflict problem by deploying different applications in the individual VMs. We can backup whole systems by performing a backup of image file of VMs.

Network configuration information

There are two different types of networks, physical and virtual networks, associated with bioinformatics platform. A physical network is a network of physical machines that are connected each other to send and receive data. The Xen VM hosts on a physical machine. A virtual network is a network that connects virtual machines logically to each other on the same physical machine. The physical Ethernet adapter bridges a virtual network and a physical network. In our bioinformatics platform, the master node of private HPC (warriors.sc.clemson.edu) and the remote backup server use the physical network. The CMD web server, BCI web server and database server use the virtual network with the identifiable public IP address and host name. Table 3.2 shows the detailed network configuration information for our bioinformatics platform.

Table 3.2 The network configuration information

Host Name Domain	Description	IP Address	MAC Address	Storage
warriors.cs.clemson.edu	The Private HPC, hosts all Xen VMs	130.127.48.117	00:1A:92:69:49:9A	**OS** /dev/md1 **Data** /dev/mapper/vg0-data
vmweb1.cs.clemson.edu	CMD web server www / dev website	130.127.48.247	00:16:36:21:f4:da	**OS** /dev/vg0/vm_cmdweb_os **Websites** /dev/vg0/vm_cmdweb_sites **Swap** /dev/vg0/cmdweb_swap
labweb1.cs.clemson.edu	BCI web server www/dev website	130.127.48.10	00:16:36:35:67:B6	**OS** /dev/vg0/vm_labweb_os **Websites** /dev/vg0/vm_labweb_sites **Swap** /dev/vg0/labweb_swap
databases	Databases server MySQL PostgreSQL	130.127.49.200	00:16:3E:7F:68:3C	**OS** /dev/vg0/vm_databases_os **Database** /dev/vg0/vm_databases_data **Swap** /dev/vg0/vm_databases_swap
roc-desktop	Remote backup server	130.127.49.163	00:22:19:1e:4e:ba	**OS** /dev/sdb1 **Backup** /dev/mapper/backup-lv_backup_01
user.palmetto.clemson.edu	HPC login node	130.127.160.100		**Job working directory** /home/pxuan/cmd

3.7 Backup

It is a significant challenge to automatically back up everything of our bioinformatics platform without stopping services. An effective and reliable backup strategy is important as well as essential since there is not a full-time system administrator to maintain the whole system, and all system services need to stay live without interruption.

To meet the requirement of this situation, a robust, efficient and reliable backup scheme is designed and implemented to support bioinformatics platform. It provides an OS-level backup mechanism to make snapshot for each VM, and then transfers image files of VMs to the remote storage server. In our design, we select rsnapshot [38] as the backup tool, and use rsync [40] utility to synchronize all backups. The usage of disk space is only the basic space of one full backup plus incrementals. Because rsnapshot only keeps a fixed number of snapshots, the amount of disk space used will not continuously grow. Figure 3.15 illustrates the backup schema of CMD virtualization platform. The host domain 0 manages the physical storage as a Physical Volume (PV) by LVM file system. Each VM has its own Logical Volume (LV) partition from the PV. For VMs in CMD platform, this results in a single LV partition hosed on LVM. When rsnapshot is configured to back up one of these LVs, it runs through the following steps:

a) Create temporary snapshots of VMs by their LVs
b) Mount snapshots in a temporary directory
c) Rotate the backup directory of the remote backup machine, making room for the current backup
d) Rsync the snapshot of VMs into the remote backup location.
e) Unmount snapshots of VMs
f) Remove the temporary snapshots of VMs

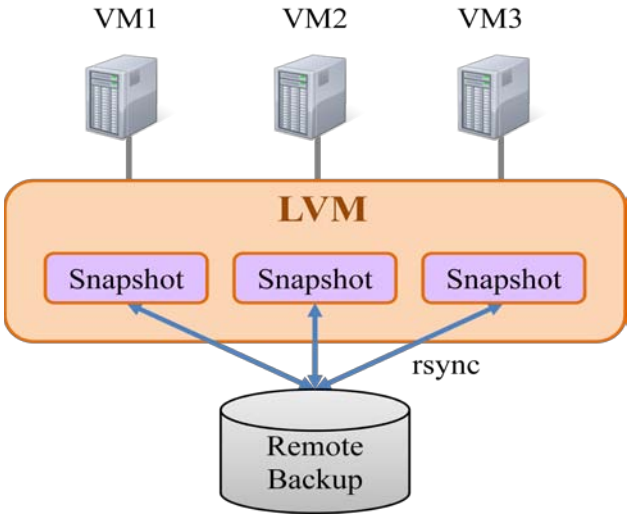


Figure 3.15 Snapshot and rsync backup solution to CMD virtualization platform

The current backup strategy is based on one week of daily backups (a rolling 7 days) and a manually static full backup. The structure of backup directory on the remote backup server is shown in Figure 3.16.

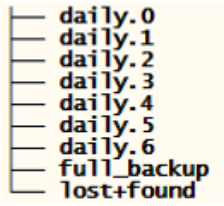


Figure 3.16 The structure of backup directory on the remote backup server

Chapter 4

Cotton Marker Database

4.1 Introduction

Cotton has had a long history as an agriculturally and industrially important crop. To improve understanding of the biological principles controlling various traits of cotton and to enhance the economic competitiveness of cotton cultivars, large-scale genetic and genomic studies are underway by cotton research groups worldwide [41-43]. To make further improvement of cotton, a large amount of molecular markers have been employed to study the tetraploid genome of cultivated cotton. In 2004, Clemson University, in association with Cotton Incorporated, launched the Cotton Marker Database (CMD), a public database and website that provides an easy-access to the publicly available single nucleotide polymorphism (SNP) and SSR markers [44].

SSRs and SNPs hosed in the CMD have been developed by many research groups all over the world within the international cotton community. In collaboration with the principal investigators of the cotton marker development projects, we have annotated the CMD data by arranging, analyzing, integrating and refining the data with an efficient interface for user access. In the following sections, we describe the current CMD database updates with the new enhancements including: new SSRs, redundancy information of SSRs, new trait/QTL feature, extensive genetic maps, new SNP data, new database design and structure, and enhanced web-based and community resources.

4.2 New SSR Projects

Compared to the previously reported number of SSR markers available through CMD [44], the current total number of cotton microsatellites has significantly increased from 3,452 SSRs in January of 2006 to 17,448 SSRs (including 312 SSR-containing RFLPs) by July of 2011. The 192 SSRs in the STV project were derived from multiple tissues of *Gossypium hirsutum*, including 150 primer pairs screened on the CMD panel [45]. The 2,937 SSR markers from the new MON project were provided by the Monsanto Company [46]. Bioinformatics analysis of the MON SSR sequences and primer pairs in comparison with the cotton SSR sequences already present in public databases revealed that these SSR primer pairs and target genomic sequences are unique and amplify about 4,000 unique marker loci in a tetraploid cotton genome depending on the germplasm analyzed [46]. Another new SSR project, DPL, contributed by the Delta and Pine Land Company, includes 200 microsatellites developed from *G. hirsutum* small insert genomic libraries enriched with multiple microsatellite motifs. Seven hundred SSRs from the new Gh project were initially evaluated for internal structure and potential for homodimer and heterodimer formation using publicly available web-based applications [47].

The HAU SSR project including 3,382 microsatellites was developed at the Huazhong Agricultural University. HAU001-HAU119 markers were developed and evaluated from 98 unique ESTs from the cDNA library of 2 – 25 day post anthesis (DPA) developing fibers from *G. barbadense* cv. Pima 3-79. Markers HAU120-HAU205 were developed from *G. barbadense* cv. Pima 3-79 using two different approaches: (i) cloning of ISSR amplified fragments and (ii) amplification using degenerate primers. A total of

1,831 new EST-SSRs were developed from the assembled cotton ESTs in the TIGR database (http://www.tigr.org): 346 from *G. arboretum* (HAU231-HAU576), 293 from *G. raimondii* (HAU577-HAU869), and 1192 from *G. hirsutum* (HAU870-HAU2061); 1047 unique EST-SSRs (HAU2062-HAU3108) were developed from ESTs released by Yuxian Zhu; 299 novel EST-SSRs (HAU3109-HAU3407) were developed from ESTs from developing fiber of *G. barbadense* acc. 3-79 [48, 49].

In addition, 2,233 markers from the NBRI SSR project (263 genomic SSRs and 1,970 EST-SSRs) were developed using four genomic libraries microsatellite-enriched for CAn, GAn, AAGn and ATGn repeats, as well as the transcriptome sequencing of five cDNA libraries of *Gossypium herbaceum*. Lastly, 312 SSR-containing RFLPs were included for the PGML project [50] .

All new SSRs were incorporated into the existing CMD data display structure, with the SSR Project pages, View and Search pages, updated Homology and Downloads pages. On the SSR View Page, the marker name is linked to a detail page including marker info, sequence type, location of longest ORF and repeat motifs in the sequence. The FASTA and BLAST server databases were updated with all the new SSR sequences from different projects as separate databases, which allow CMD users to perform a batch query using a specific SSR project. In addition, three major protein databases used in the BLAST and FASTA servers in the Tools section were updated with the latest versions: UniProtKB/Swiss-Prot (release 2011.08), UniProtKB/TrEMBL (release 2011.08), and TAIR Arabidopsis (release 2010.12). Homology searches were performed for 1712,448

SSR sequences using all three updated protein databases. Recent homology search results are available on the CMD Downloads page (Homology Data section).

4.3 Primer Redundancy Information of SSRs

Currently, 18,002 CMD microsatellite primer sequences have been checked using the CMD primer redundancy analysis tool. Any of the following criteria were considered as redundant: (i) identical primer pairs; (ii) identical forward primers; (iii) identical reverse primers; (iv) forward primer identical to reverse and vice versa. From this analysis, 85.7% of the microsatellite primer sequences checked were considered to be unique and noted accordingly in the database. The Primer redundancy analysis tool is a Perl package built around the pair wise comparison algorithm to create clusters of identical sequences in accordance to the threshold value specified for the analysis. The primer redundancy threshold value is the "global sequence identity" calculated as number of identical nucleotides in alignment divided by the full length of the shorter sequence. Based on the input threshold value, we construct clusters of sequences and search for identity. A threshold of 0.81 (81%) was chosen for the analysis after plotting the threshold values from 70% to 100% against the redundant primer counts. Table 4.1 gives the summary of primer redundancy to the threshold 81%. In Figure 4.1, the redundant primer counts are obtained for each of the individual threshold values.

In the first quarter of 2009 the new feature Primer Redundancy was added to the View section. In the third quarter the section was updated with the new primer redundancy results. The new results were obtained after performing further data analysis based on the presence of the combined redundancy of the SSR primer sequences and

repeat motifs. MON SSR project data was not included in the primer redundancy analysis due to the lack of data on the target repeat motifs.

Table 4.1 The summary of primer redundancy processing.

Number of primer sequences analyzed (forward and reverse)	18,002
Total number of redundant primer sequences found	2,570
Total number of non-redundant primer sequences	15,432
Threshold value for the analysis	81
Total number of redundant primer pairs	1,422
Total number of completely matched sequence pairs (match percentage 100)	940
Total number of closely matched sequence pairs (match percentage 90-99)	280
Total number of closely matched sequence pairs (match percentage 81-89)	202
Type of primer sequence match (pairs)	
Forward-forward match	460
Reverse-reverse sequence match	414
Forward-reverse sequence match	232
Reverse-forward sequence match	316
Both forward-forward and reverse-reverse match	103

Note: data presented in the table was retrieved from the CMD website.

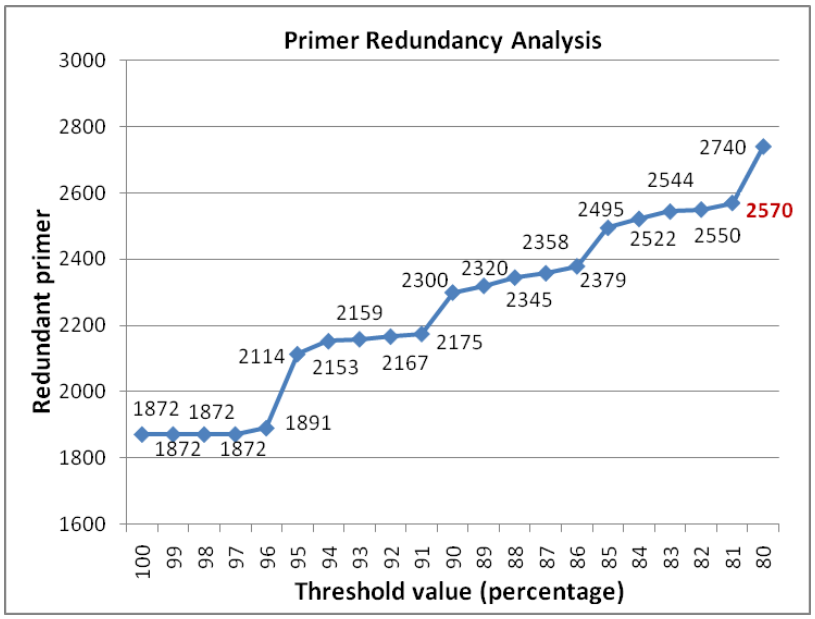


Figure 4.1 The redundant primer counts obtained for each of the individual threshold values

4.4 SSR Redundancy Detection using SVM Machine Learning Approach

Microsatellites or SSRs are used as molecular markers with wide-ranging applications in the field of cotton molecular breeding. CMD provides centralized access to publicly available cotton molecular data. In collaboration with the contributing researchers, we have summarized and provided high quality data for 17,488 SSRs displayed through CMD. However, SSR redundancy is common and inevitable issue for projects coming from different research groups. The method of SSR redundancy detection using the SSR-containing sequence alignment approach gives high number of false-positives even when applying stringent parameters, since the similarity identification is based only on the sequence comparison. To improve the accuracy of the redundant SSRs detection and to reduce the cost of expert intervention, we propose the application of machine learning based on the SVM machine learning approach.

4.4.1 Materials and Methods

We choose LIBSVM program [51] as machine learning program because it has efficient multi-class classification, and we use cross validation method. We weight SVM for unbalanced data. Specially, LIBSVM can automatically select the model that can generate the contour of greatest cross valuation accuracy. This feature lets us more easily evaluate and select parameters for our SVM program. Figure 4.2 presents the workflow of SVM filtration with three phases: generate SVM model via the training data, verify the

performance via the testing data, and predict/refine SSR redundancy based on the result of sequence similarity.

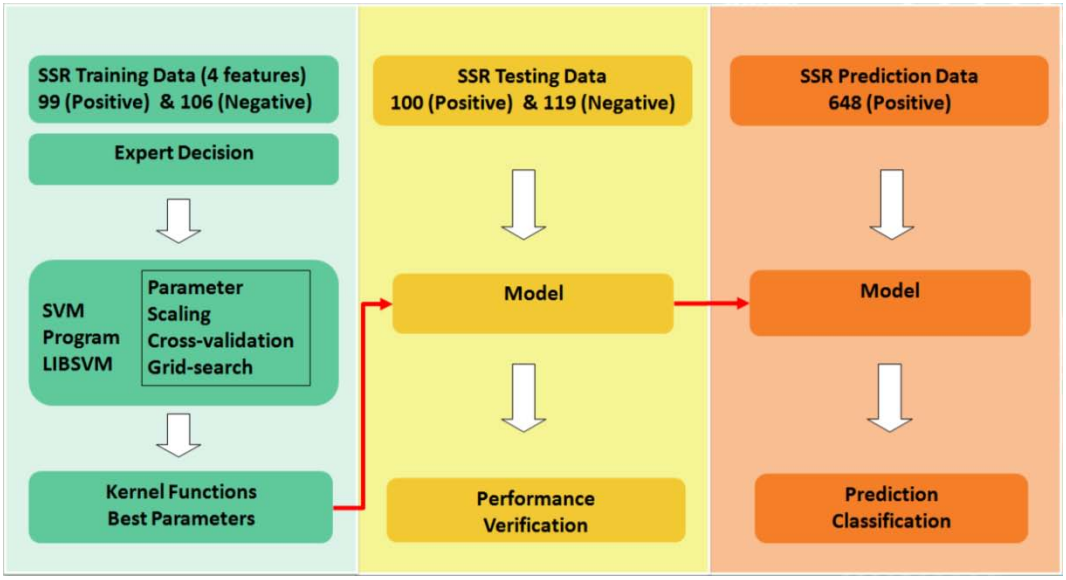


Figure 4.2 The SVM machine learning workflow.

The feature selection for machine learning is a critical and also challenging task, which usually requires an iterative approach. For our current problem we select a set of five features related to properties of SSRs which are likely to influence a human expert when classifying an SSR redundancy, including: percentage match for primer redundancy, primer match type, motif similarity, percentage match for SSR sequence and map position in the genetic map. The first four SSR features are selected for the machine learning approach, and the last feature is used to help the expert to verify the predicted result. The CMD SSR dataset (847 markers) is used as training, testing and prediction sets for the SVM algorithm. Table 4.2 describes the algorithm for the training data selection.

Percent match of primer sequences: The SSR primer sequence is an important referenced factor in genetic research. It is used to isolate targeted sections of DNA for amplification in PCR. The primer sequence alignment can be calculated by CD-HIT program [52].

Primer match type: Type 1 - forward to forward match, and reverse to reverse match; Type 2 - forward to reverse match, or reverse to reverse match.

Motif similarity: SSR motif similarity is another important factor reflecting the degree of SSR redundancy.

Percent match of SSR-containing sequences: BLAST search allows the comparison for a pair of SSRs, and identify them above a certain threshold.

SSR genetic map position: Based on this feature, the training data were manually selected and the final results were evaluated.

4.4.2 Result

The SVM approach with different kernel functions is applied to develop an accurate model for SSR redundancy detection. In our experiment, we select four different Kernel Functions to compare the performance of SVM based on optimal parameters C = 512 and gamma = 0.0078125. The ROC curve analysis which is generated based on the cross validation results (Figure 4.3) indicate the remarkable performances of the SVM approach. The high accuracy and F-score in Table 4.3 shows that the SVM-based machine learning method can identify SSR redundancy with the high predictive performance.

Table 4.2 The algorithm to select the SSR redundancy training data.

No.	Condition	Action
Case 1	No mapping data for any marker.	OMIT
Case 2	Two markers present once on the same chromosome of same map.	PROCEED
Case 3	Two markers mapped once on the same chromosome of two different genetic maps, two chromosome bridged by 0 or 1 marker.	OMIT
Case 3 (A)	Either of the two markers present more than once on same chromosome of same map (one of the 2 or both may provide clue for redundancy).	OMIT
Case 4	Two markers mapped once on the same chromosome of two different genetic maps, two chromosome bridged by at least 2 flanking markers each mapped once.	PROCEED
Case 4 (A)	Two markers mapped once on same chromosome of two different maps, but among flanking bridged markers one being in paralog duplication	PROCEED

Note: The four cases presented above are exclusive to each other. The process may be iterative. Once a first pair of markers has been validated (e.g. marker 1 and marker2 are synonyms), they may become informative in terms of additional bridges for others: for the next pair of markers formerly under case 3 this new information may result in a case 4 situation.

Table 4.3 Evaluation of results obtained for the tested data.

Kernel Function	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
Linear	98	2	117	2	98%	98.32%	98.00%	98.17%	98.00
Polynomial	0	0	119	100	0%	100.00%	--	54.33%	--
Radial Basis	97	2	117	3	97%	98.32%	97.98%	97.71%	97.31
sigmoid	99	3	116	1	99%	97.48%	97.06%	98.17%	98.02

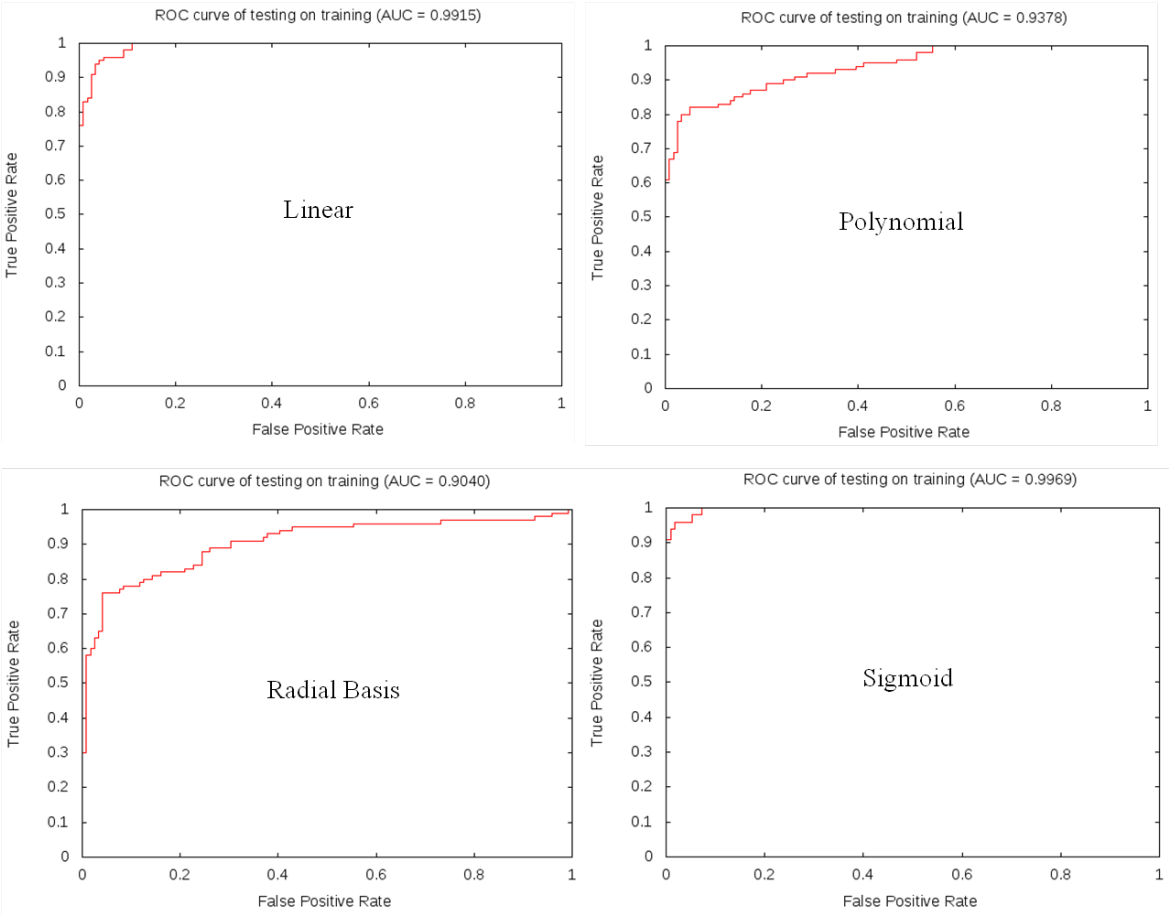


Figure 4.3 ROC curve analysis

4.5 QTL/Traits Feature and Cotton Genetic Maps

In breeding, cotton cultivars have been grown and cross-bred to produce desirable, agronomically important traits, which are very often associated with combinations of several genes, called Quantitative Trait Loci (QTLs). To better understand and determine where particular traits of interest are located on chromosomes of a specific species of cotton, researchers have identified a large number of DNA molecular markers linked to certain traits and corresponding QTLs. These molecular markers include, but are not limited to, SNPs and SSRs, or microsatellites.

The total number of agriculturally important cotton traits displayed through CMD is currently 44, which corresponds to 76 trait symbols and 884 mapped QTL positions on 14 cotton genetic maps. The QTL information has been uploaded into the CMD Comparative Map Viewer (CMap) accessible from the CMD Homepage.

In the past four years, 23 new cotton genetic map sets (21 genetic maps corresponding to individual crosses, 1 consensus map and 1 reference map), were added to the previously available 4 genetic maps, were added to the CMap Viewer [53] on the CMD website. In addition, we annotated the CMD cotton traits with the related information about the trait-associated genetic markers from the cotton genetic maps and represented their associations through the CMD-CMap connection. Specifically, the traits are annotated and linked in CMD with the trait description, aliases, published symbol(s), QTL/gene name(s), associated CMD marker(s), cross name(s), marker interval for QTLs, R-square value, genetic linkage group information, genetic map positions and corresponding publications.

Chapter 5

Quantitative Structure-activity Relationship (QSAR) Study on Glycan Array Data

5.1 Introduction

Glycan-binding proteins play critical roles in many physiological and pathological processes [54], including inflammation and cancer [55-57], growth and development [58-60] and microbial pathogenesis [61-64]. In order to understand the biology of glycan-binding proteins, it is essential to identify their glycan-binding specificities. Recently, glycan array technology [65-68] provided a high throughput method to simultaneously measure the binding levels of a certain glycan-binding protein to a large number of glycan molecules. The newest version (V5.0) of the glycan array from the Consortium for Functional Glycomics (CFG) [66] contains 611 glycan chains. Currently, large amounts of glycan array data are freely available on the CFG website (www.functionalglycomics.org), and this number is still increasing. These glycan array data have opened up opportunities to discern the binding specificities for glycan-binding proteins.

The glycan array data usually are very complex, and simple visual inspections may not be able to identify the binding specificities of glycan-binding proteins. This poses a great challenge to extract binding specificities of glycan-binding proteins from glycan array data [69]. Recently, Porter et al. (2010) proposed motif-based methods to discern the sub-structures that contribute to the binding intensities of glycan array to a

specific glycan-binding protein. Porter et al. manually generated a list of 63 motifs that are sub-structures of glycan chains identified previously by biological experiments. By comparing the enrichment of those motifs in high intensity and low intensity data (intensity segregation) or by statistical testing between glycan data with a certain motif and glycan data without a certain motif (motif segregation), Porter et al. (2010) were able to find motifs that represent binding specificities. However, such predefined motifs may not be sufficient to identify all glycan binding specificities.

We have developed a novel quantitative structure-activity relationship (QSAR) method to analyze glycan array data. First, we automatically generated different size sub-trees from glycan chains as our features. Then, we established the relationship between sub-tree features and glycan array data using the PLS regression. We demonstrated our QSAR method on glycan array data of different glycan-binding proteins. We were able to identify sub-trees that represent the glycan binding specificities of glycan-binding proteins using the regression coefficients of PLS regression. We also showed that the sub-tree features may be better representations of the glycan binding specificity than the motifs defined by Porter et al. (2010) are. Furthermore, we developed a user-friendly web tool to facilitate the rapid and automatic analysis of glycan array data. A complete description of our results and methods is given in the sections below.

5.2 Results

5.2.1 Coding Glycans using Sub-tree Features

Glycan chains consist of different kinds of saccharides, such as glucose (Glc), fucose (Fuc), galactose (Gal), N-acetylglucosamine (GlcNAc), mannose (Man), and N-acetylgalactosamine (GalNAc). The structure of a glycan chain can be represented as a rooted tree. Figure 5.1 shows an example glycan chain that consists of five different saccharides. The binding specificity of a glycan chain to glycan-binding protein usually relies only on its substructures [64, 69]. In order to capture the structure characteristics of glycan chain, we parsed the glycan tree into four sets of sub-trees [70, 71], each of which has mono-, di-, tri-, or tetra-saccharide sub-tree respectively. Figure 5.2 shows that the example glycan chain in Figure 5.1 has been decomposed into five mono-saccharide sub-trees, five di-saccharide sub-trees, five tri-saccharide sub-trees, and four tetra-saccharide sub-trees, respectively. For each version of glycan array from the Consortium for Functional Glycomics (CFG) [66], we generated four sets of sub-trees for the glycan on the array, including mono-, di-, tri-, and tetra-saccharide sub-tree sets. For example, the CFG glycan array version 2.0 contains 264 glycan chains. We obtained 112 mono-saccharide sub-trees, 280 di-saccharide sub-trees, 385 tri-saccharide sub-trees and 318 tetra-saccharide sub-trees from those 264 glycan chains. The four sets of sub-trees for the CFG glycan array version 2.0 are listed in Appendix A Table A-(I-IV).

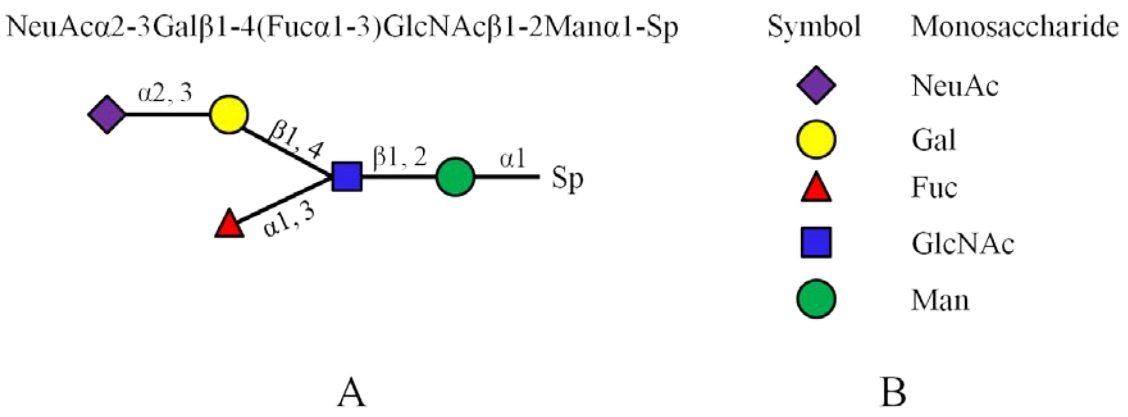


Figure 5.1 An example of glycan chain and its structure. The Sp denotes the spacer arm attached to array.

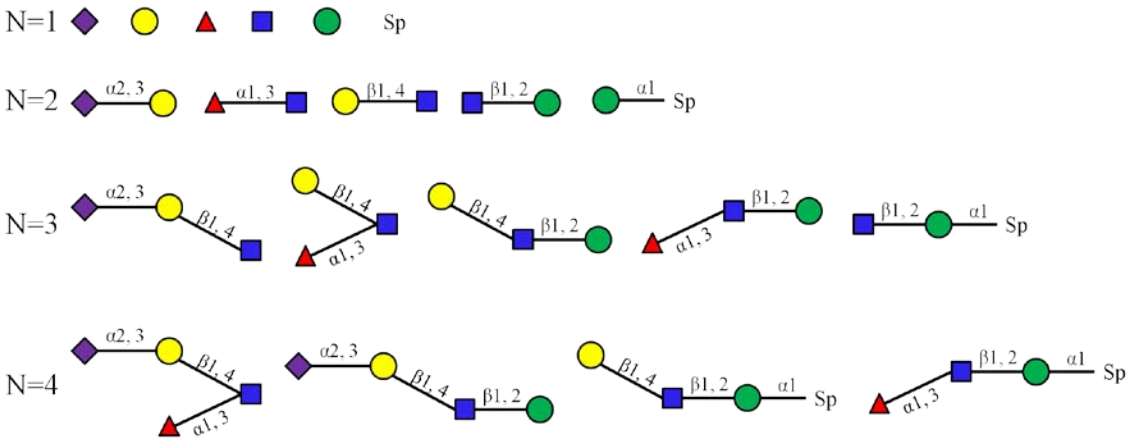


Figure 5.2 An example of decomposing the glycan chain in Figure 5.1 into different sub-trees. The N indicates the number of saccharide in each sub-tree.

In order to represent better the sub-structural characteristics, we included also the bond information in the sub-trees. For each saccharide, we included the positions of its bond connections in its representation. Different mono-saccharide sub-trees will be generated if the same saccharide has different bond connections. For example, we had five mono-saccharide sub-trees for galactose (Gal): (2, 3Galβ); (2, 3Galβ1); (2, 4Galβ1); (2Galβ) and (2Galβ1) from the glycan chains of CFG version 2.0 array. Furthermore, the

di-saccharides are also represented differently if the bonds between the same pair of saccharides are different. For example, we had two di-saccharide sub-trees between N-acetylglucosamine (GlcNAc) and galactose (Gal): (3,4GlcNAcβ1-3Galβ1) and (3,4GlcNAcβ1-4Galβ1). With the bond information, the sub-trees extracted from glycan chains can help distinguish the glycan chains structurally to a certain extent.

After obtaining the sub-trees, we used them as features to code the structures of glycan chains on the glycan array. This new coding system has an advantage over the motif-based approach in which the sub-tree features are more precise and more flexible. Many sub-structures potentially cannot be represented well in motif-based features since there have been only 63 defined motif features (Porter, A., et al. 2010). In contrast, the number of sub-tree features in our method can be much larger (e.g. 112 mono-saccharide sub-trees, 280 di-saccharide sub-trees, 385 tri-saccharide sub-trees and 318 tetra-saccharide sub-trees for the CFG glycan array version 2.0). Our method requires more computation than the motif-based method, but this fact shall not be considered a significant limitation of our method.

The structures of glycan chains were obtained from the CFG website. The CFG glycan array version 2.0 data of plant lectins were also downloaded from the CFG website. Table 5.2 list all 52 plant lectins that we analyzed.

For each glycan chain in a certain version of CFG glycan array, we coded one vector based on each set of mono-, di-, tri-, or tetra-saccharide sub-trees. The elements in each vector were 1 and 0. If a glycan chain contains a sub-tree, we coded the feature with 1; otherwise, we coded the feature with 0. Then, feature vectors were used for PLS

regression study. A Java program was implemented to automatically parse the glycan chains into mono-, di-, tri-, or tetra-saccharide sub-trees, and then code the glycan chains with different sub-trees.

5.2.2 PLS Regression on Glycan Array Data using Different Features

We first applied the PLS regression to the glycan array data of three plant lectins: Concanavalin A (ConA), Vicia Villosa Lectin (VVL) and Wheat Germ Agglutinin (WGA), which were also studied by motif-based methods [69]. The binding specificities of ConA and VVL are relatively simple. A visual inspection may help to identify some common features from the data. For example, it is shown clearly from the data that ConA binds to the glycans that contain terminal N-acetylglucosamine (GlcNAc). On the other hand, the binding specificity of WGA is broad and cannot be determined easily by visual examination. To understand how different sub-structures contribute to binding specificity, we performed PLS regression studies on the glycan array data of those three plant lectins using the mono-, di-, tri-, and tetra-saccharide sub-tree features as well as the motif features of Porter et al. (2001). We first examined the percentage of variances of binding intensities that can be explained using PLS regression models. The percentage of variance explained measures the amount of variation in the given data that a regression model accounts for and it can be used to indicate how well the regression model is. The higher the percentage of variance explained is, the better the PLS regressions perform and the better the sub-tree features are. Figure 5.3 plots the percentage of variance explained in the binding intensities of three plant lectins against the number of latent variables

(components) in PLS regression. The number of components is automatically determined by their contributions to the variance (see Method section for more details). Thus, the number of components varied for PLS regressions using different features. Figure 5.3Figure 5.3 shows that the PLS regression using di-saccharide sub-trees achieved the highest percentage of variance explained for all three glycan array data. The PLS regression using mono-saccharide sub-trees achieved high percentage of variance explained in glycan array of ConA and the PLS regression using tri-saccharide sub-trees achieved high percentage of variance explained in glycan array of WGA. The PLS regression using tetra-saccharide sub-tree and motif features did not obtain high percentage of variance explained in all three glycan array data. Thus, the tetra-saccharide sub-tree and motif features cannot fully capture the intensity variations in those glycan array data. These results implied that the motif-based method may not have sufficient sensitivity to cover all binding specific sub-structures.

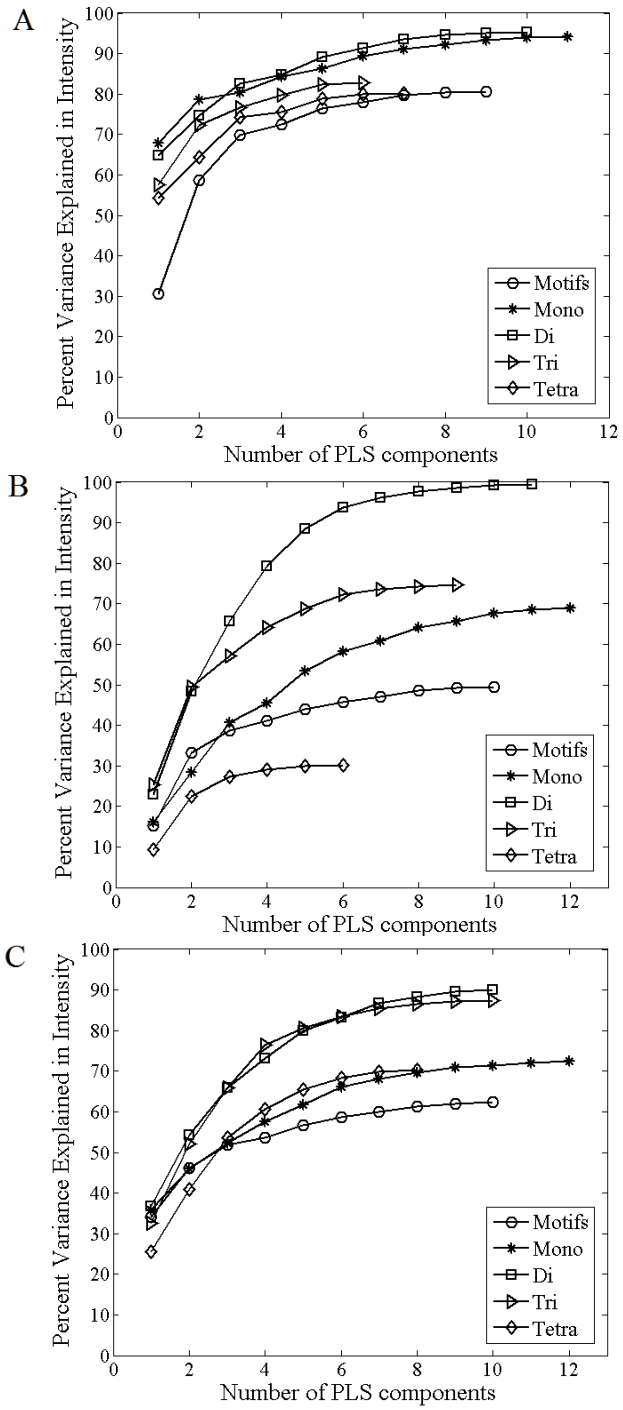


Figure 5.3 Plot of the percentage of variance explained in the binding intensities of glycan array data of three plant lectins against the number of components in PLS. Four sub-tree features and the motif features of Porter et al. [69] are used for PLS regression. A) ConA, B) VVL, C) WGA.

Then, we calculated the R^2 statistics of PLS regressions. The R^2 is a statistical measurement indicating how well a regression approximates real data. The R^2 analysis (Table 5.1) is consistent with these results of variance explained above. For ConA, the PLS regressions with all five features can obtain an $R^2>0.8$. For VVL, only the PLS regression using di-saccharide sub-trees as features can obtain a significant high $R^2=0.9955$. For WGA, the PLS regression using both di- and tri-saccharide sub-trees can obtain high $R^2>0.8$. Those results confirmed that di-saccharide sub-trees are good features for characterizing the glycan array data of three plant lectins. We also tested the PLS regression using di-saccharide sub-trees on glycan array data of more than 50 plant lectins (Table 5.2). We obtained good results ($R^2>0.8$) on most of the glycan array data except two of them, which have good regression results using tri-saccharide sub-trees as features. To further examine the results of PLS regressions, we plotted the observed intensities against the fitted intensities calculated by PLS regression using di-saccharide sub-trees for all three plant lectins. As shown in Figure 5.4, there are good correlations between the observed intensities and fitted intensities for ConA, VVL and WGA. The dots in Figure 5.4C are distributed more widely than those in Figure 5.4A and Figure 5.4B, which is consistent with relatively low R^2 value obtained by the PLS regression on glycan array data of WGA. Those plots implied that the di-saccharide sub-trees can represent the binding specificity of ConA, VVL and WGA well.

Table 5.1 The R^2 of PLS regressions on glycan array data of different glycan-binding proteins using different features.

	Motifs	**Mono**	**Di**	**Tri**	**Tetra**
ConA	0.8052	0.9428	**0.9539**	0.8276	0.8021
VVL	0.4943	0.6893	**0.9955**	0.7461	0.3024
WGA	0.6242	0.7132	**0.9002**	0.8742	0.7027
PNA	0.7774	0.5752	0.9603	**0.9966**	0.7619
SNA	0.6760	0.7871	**0.9085**	0.7431	0.6509
DC-SIGN	0.4190	0.5490	0.9179	**0.9533**	0.8521
Siglec-8	N/A	0.9882	0.9927	**0.9969**	0.9949

Note: The highest values of R squares are highlighted in bold.

Table 5.2 The highest R^2 of PLS regressions on glycan array data of different glycan-binding proteins.

Glycan-binding protein name	**Glycan-binding protein full name**	**Number of PLS component**	**R^2**	**Features Vector Used**
AAL	*Aleuria Aurantia* Lectin	9	0.941	Di
ABA	*Agaricus bisporus* agglutinin (mushroom)	10	0.8904	Di
ACL	*Amaranthus cruentus* lectin (red amaranth, purple amaranth)	13	0.9939	Di
AMA	Arum maculatum agglutinin (lords and ladies)	8	0.9022	Tri
APA	*Allium porrum* agglutinin	8	0.9189	Tri
ASA	*Allium sativum* agglutinin/lectin	13	0.9018	Di
BPL	*Bauhinia purpurea* agglutinin	10	0.937	Di
CA	*Cymbidium* agglutinin	8	0.9503	Di
CAA	*Caragana arborescens* agglutinin (Siberian pea tree)	11	0.8852	Di
ConA	Concanavalin A (*Canavalia ensiformis*, jack bean)	10	0.9539	Di
Crocus	*Crocus Vernus* Agglutinin	12	0.8017	Di
CSA	*Cytisus sessilifolius* agglutinin (Portugal broome)	13	0.9926	Di
DBA	*Dolichos biflorus* agglutinin (horse gram)	14	0.9711	Di
ECA	*Erythrina cristagalli* agglutinin (cocks comb coral	10	0.9173	Di

	tree)			
EEL	*Euonymus Europaeus* Lectin	10	0.8801	Di
GNL	Peanut nodule lectin (*Arachis hypogaea)*	17	0.9616	Di
GSL_I	*Gerardia savaglia* lectin (false foxglove)	11	0.9671	Di
GSL_II	*Gerardia savaglia* lectin (false foxglove)	9	0.9542	Di
HHL	*Hippeastrum Hybrid* Lectin	10	0.9495	Di
Jacalin_CoreD	*Artocarpus heterophyllus* (bread fruit tree)	13	0.9637	Di
Jacalin_CoreH	*Artocarpus heterophyllus* (bread fruit tree)	12	0.9424	Di
LCA	*Lens culinaris* agglutinin (lentil)	12	0.9935	Di
LEL	*Loranthus europaeus* lectin (loranthus, misteltoe)	12	0.9221	Di
LTL	*Lotus tetragonolobus* agglutinin	14	0.9157	Di
MAA	*Maackia amurensis* agglutinin/lectin	13	0.977	Di
MAL_CoreD	*Maackia amurensis* agglutinin/lectin	14	0.992	Di
MAL_CoreH	*Maackia amurensis* agglutinin/lectin	15	0.9214	Di
MPA	*Maclura pomifera* agglutinin	10	0.995	Di
MPL	*Maclura pomifera* agglutinin	12	0.9967	Di
NPL	*Narcissus pseudonarcissus* agglutinin/lectin	10	0.9436	Di
PHA_L	Leucoagglutinating isolectin of PHA	13	0.8511	Di
PMA	*Polygonatum multiflorum* lectin (common Solomon's seal)	10	0.9148	Di
PNA	*Arachis hypogaea* agglutinin (peanut)	12	0.9122	Di
PSA	*Pisum sativum* agglutinin (garden pea, common pea)	10	0.9815	Di
PTL_I	*Psophocarpus tetragonolobus* agglutinin (goa bean, winged pea)	10	0.993	Di
PTL_II	*Psophocarpus tetragonolobus* agglutinin (goa bean, winged pea)	9	0.9472	Di
RCA_I	Ricinus Communis Agglutinin I	10	0.8998	Di

Rcroc	Rhizoctonia crocorum	18	0.8209	Di
SBA	Soybean agglutinin (*Glycine max*, soya bean)	11	0.9801	Di
SJA	*Sophora japonica* agglutinin (Japanese/Chinese pagoda tree)	8	0.8408	Di
SNA	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	19	0.9072	Di
SNA_I	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	12	0.902	Di
SNA_II	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	12	0.8459	Di
SNA_III	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	11	0.9567	Di
SNA_IV	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	10	0.9782	Di
SNA_V	*Sambucus nigra* agglutinin (elderberry, eldertree, elder)	12	0.9705	Di
STL	*Solanum tuberosum* agglutinin (potato)	10	0.9617	Di
UEA_I	*Ulex europaeus* agglutinin (furze, gorse)	10	0.9149	Di
VFA	Vicia fava	11	0.9625	Di
VVL	*Vicia villosa* agglutinin (hairy vetch)	11	0.9954	Di
WFA	*Wisteria floribunda* agglutinin (Japanese wisteria)	10	0.9322	Di
WGA	Wheat germ agglutinin (*Triticum vulgare*)	10	0.8997	Di

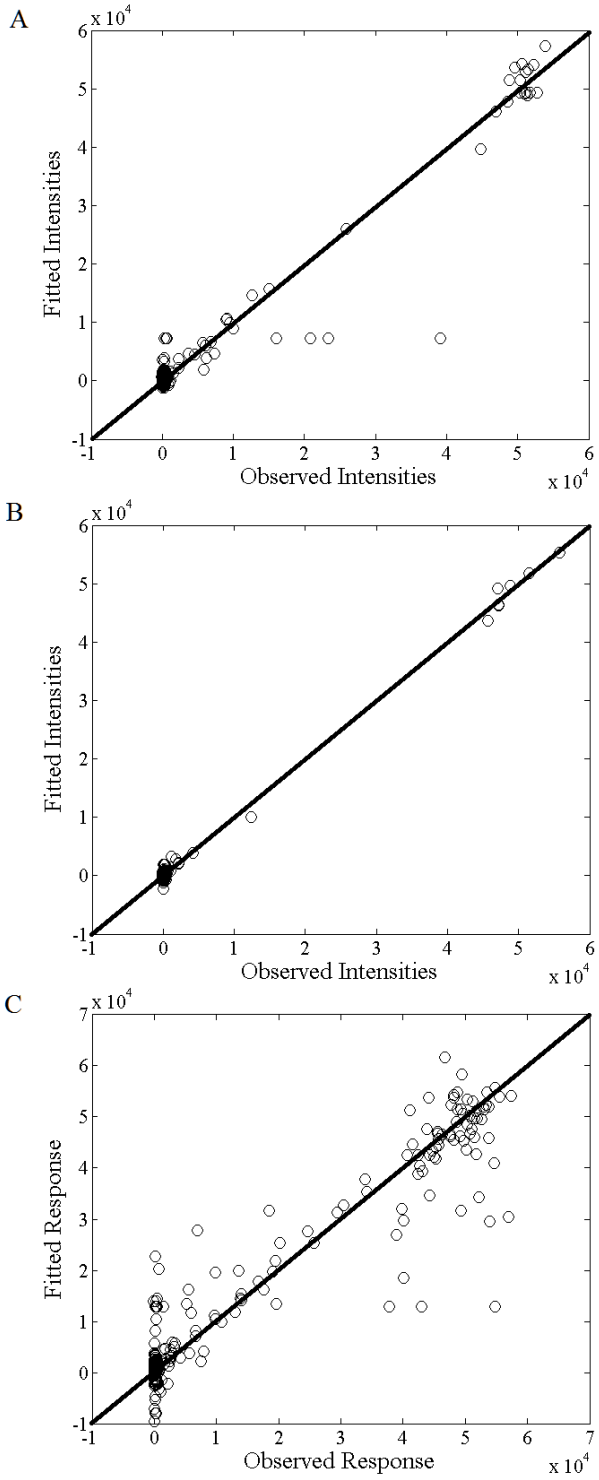


Figure 5.4 Plot of observed intensities against the fitted intensities calculated by PLS regression using di-saccharide sub-trees as features. The black lines indicate the line of y=x. A) ConA, B) VVL, C) WGA.

5.2.3 Identification of Significant Structural Features in Glycans

The PLS regression has established the relation between the response y and original predictors X as a multiple regression model:

$$y_m = \sum_k B_{mk} X_{ik} + f'_m \tag{5.1}$$

where vector f' (n×1) denote the regression errors and matrix B (p×1) denote the PLS regression coefficients and can be calculated by:

$$B_m = \sum_i C_{mi} W_{ki} \tag{5.2}$$

Then, the significant predictors can be selected based on the values of regression coefficients from PLS regression, which is called the PLS-Beta method [72].

To select the significant sub-trees using regression coefficients, we modeled the distribution of regression coefficients as a Gaussian distribution. The plots of the regression coefficient distribution obtained from PLS regression on three plant lectins showed that approximating the coefficient distribution as Gaussian distributions is reasonable (Figure 5.5, Figure 5.6 and Figure 5.7). Then, we determined the significant regression coefficients based on z-score: $z = \frac{B_i - u}{\sigma}$, where u is the average of regression coefficients and σ is the standard deviation of regression coefficients. We first calculated the z-score value for each coefficient. We selected the sub-trees whose regression coefficients with z-score larger than a threshold. The higher the z-score, the less number of sub-trees are selected.

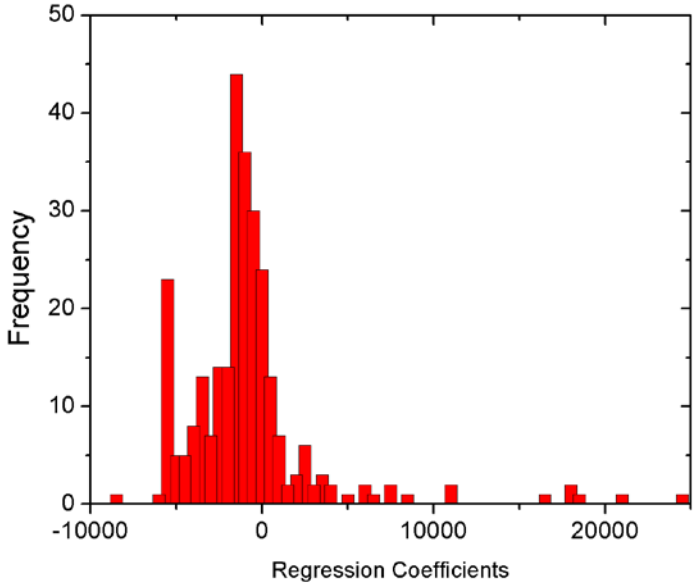


Figure 5.5 The distribution of PLS regression coefficients (Beta values) obtained from glycan array data of ConA.

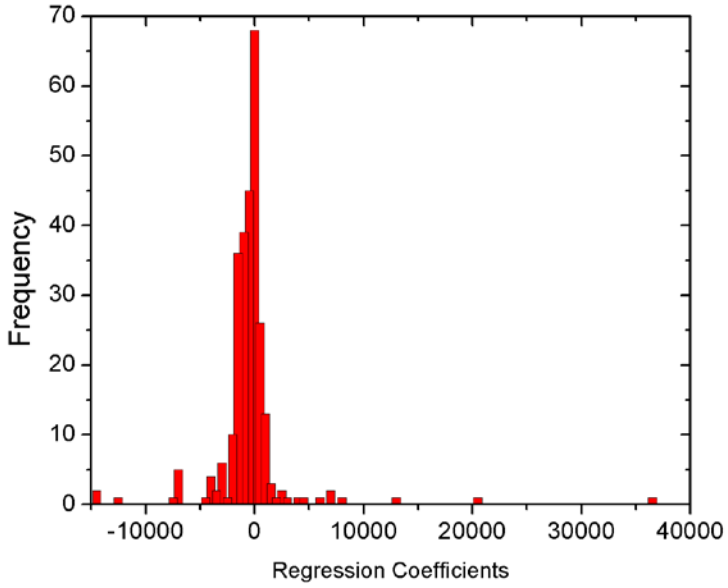


Figure 5.6 The distribution of PLS regression coefficients (Beta values) obtained from glycan array data of VVL.

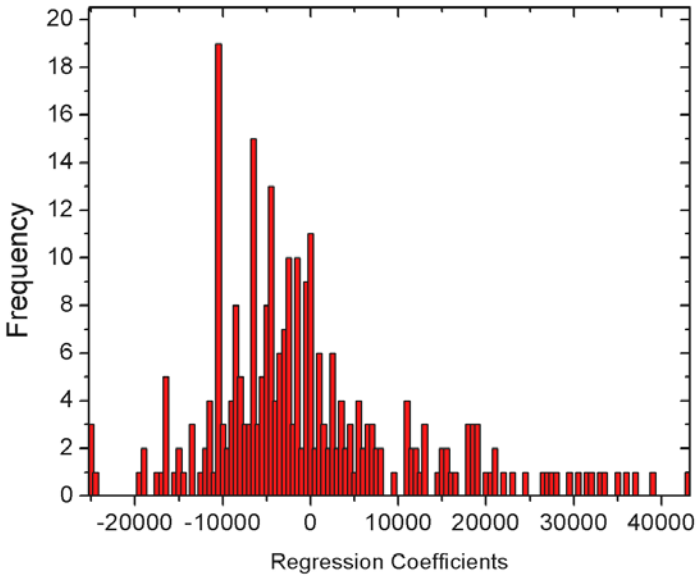


Figure 5.7 The distribution of PLS regression coefficients (Beta values) obtained from glycan array data of WGA.

We applied the PLS-Beta method (see Methods section for more details) to identify significant sub-trees from the PLS regressions of glycan array data. Table 5.3, Table 5.4 and Table 5.5 list the significant di-saccharide sub-trees binding to three plant lectins. A significant positive coefficient value indicates the corresponding sub-trees have high binding intensities whereas a negative coefficient value suggests the existence of the sub-tree feature will reduce the binding intensity. A negative co-efficient value can be achieved when two glycan chains contain the same sub-tree structure, one with high binding intensity and the other with low binding intensity. For ConA, we identified nine di-saccharide sub-trees, which cover all 19 glycan chains (Table 5.3) with high binding intensities. Among these nine di-saccharide sub-trees, alpha-linked mannose (Man) is involved in seven, and four of these alpha-linked mannoses locate at terminal. Both motif and intensity segregation methods [69] ranked the “terminal Mannoseα” as a

significant motif. The QSAR results showed internal alpha-linked mannose may also contribute to the binding specificity of ConA as four of our significant di-saccharides contained internal alpha-linked mannose. The QSAR method identified that a di-saccharide: 3,6Manβ1-4GlcNAcβ1, contributed to binding of 10 glycan chains in CFG glycan array version 2.0:

- 50 (Manα1-3(Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Gly)
- 51 (GlcNAcβ1-2Manα1-3(GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Gly)
- 52 (Galβ1-4GlcNAcβ1-2Manα1-3(Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Gly)
- 53 (Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-3(Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Gly)
- 54 (Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-3(Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Sp8)
- 192 (Manα1-6(Manα1-2Manα1-3)Manα1-6(Manα1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn), 193 (Manα1-2Manα1-6(Manα1-3)Manα1-6(Manα1-2Manα1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn)
- 194 (Manα1-2Manα1-2Manα1-3(Manα1-2Manα1-3(Manα1-2Manα1-6)Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn)
- 197 (Manα1-6(Manα1-3)Manα1-6(Manα1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn)

- 198 (Manα1-6(Manα1-3)Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn)

This di-saccharide is a subset of both "N-glycan high mannose" and "N-Glycan complex" that are identified as significant motifs by both motif and intensity segregation methods [69]. The "N-glycan high mannose" motif is defined by Porter et al. (2010) as "a glycan chain with a Manα1-3(Manα1-6(Manα1-3)Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ base" and the "N-Glycan complex" is defined as "a glycan chain with a GlcNAcβ1-2Manα1-3(GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ base" In CFG glycan array version 2.0, the "N-glycan high mannose" motif exists in five glycan chains: 192, 193, 194, 197, 198 and the "N-Glycan complex" motif also exists in five glycan chains: 51, 52, 53, 54 and 201 [69]. The glycan chain 201 (Neu5Acα2-3(Galβ1-3GalNAcβ1-4)Galβ1-4Glcβ-Sp0) does not contain the di-saccharide 3,6Manβ1-4GlcNAcβ1 and its binding intensity with ConA is low. We then performed similar motif segregation study to compare the di-saccharide 3,6Manβ1-4GlcNAcβ1 with the "N-glycan high mannose" and "N-Glycan complex" motifs. We used the two-tail unpaired t-test. The p-values are 2.45E-21 for the "N-glycan high mannose" and 1.18E-15 for the "N-Glycan complex" without counting glycan chain 201. On the other hand, the p-value for 3,6Manβ1-4GlcNAcβ1 is 4.61E-33. Thus, the di-saccharide 3,6Manβ1-4GlcNAcβ1 may be a better representation of the binding specificity of those 10 glycan chains. We also identified a significant di-saccharide, Glcα1-4Glcβ, which corresponds to the terminal glucose motifs identified by motif-based methods [69]. Furthermore, the listing of glycan chains that contain the significant di-saccharide in Table 5.3 shows that some

of those significant di-saccharides are dependent of each other. For example, 2Manα1-3Manα and 3Manα-Sp9 both exist in glycan chains 189 (Manα1-2Manα1-2Manα1-3Manα-Sp9) and 191 (Manα1-2Manα1-3Manα-Sp9). We may be able to merge them as a tri-saccharide: 2Manα1-3Manα-Sp9.

Table 5.3 The significant di-saccharide sub-trees binding specifically to ConA.

di-saccharide	Regression Coefficient	Glycan numbers
Man5-Asn	46769.84	199
α-D-Man-Sp8	46010.46	9
3,6Manα-Sp9	24403.98	190, 195, 196
Glcα1-4Glcβ	20758.13	177
3,6Manβ1-4GlcNAcβ1	18295.47	50, 51, 52, 53, 54, 192, 193, 194, 197, 198
2Manα1-3Manα	17959.44	189, 191
3Manα-Sp9	17959.44	189, 191
Manα1-3,6Manα	16141.64	195, 196
Manα1-3,6Manβ1	10828.64	50, 198

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant di-saccharide in CFG glycan array V2.0 are listed. The regression coefficients are obtained by PLS regression using di-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.0.

Table 5.4 The significant di-saccharide sub-trees binding specifically to VVL.

di-saccharide	RegressionCoefficient	Glycan numbers
GalNAcα1-3Galβ	47501.78	86
GalNAcβ1-4GlcNAcβ	46390.05	92, 93
a-GalNAc-Sp8	44524.76	10
b-GalNAc-Sp8	44432.05	20
GalNAcβ1-2,3Galβ	36068.33	89
GalNAcβ1-3Galα1	20391.86	90
Galα1-2,3Galβ	-14522	99
GalNAcα1-2,3Galβ	-14717.1	84

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant di-saccharide in CFG glycan array V2.0 are listed. The regression coefficients are obtained by PLS regression using di-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.0.

Table 5.5 The significant di-saccharide sub-trees binding specifically to WGA.

di-saccharide	Regression Coefficient	Glycan numbers
GalNAcα1-3Galβ	42672.23	86
(6OSO3)GlcNAcβ-Sp8	38904.61	47
GlcNAcβ1-4MDPLys	36691.11	168
GalNAcβ1-3,4GlcNAcβ	35613.5	91
GlcNAcβ1-3,6GlcNAcα	34575.78	121, 159
β-GlcNAc-Sp0	33127.88	21
GalNAcα1-2,3Galβ	32604.2	84
GlcNAcβ1-6Galβ1	31649.8	176
α-GalNAc-Sp8	31345.39	10
β-GlcNAc-Sp8	30441.15	22
GalNAcβ1-4GlcNAcβ	29082.38	92, 93
GlcNAcα1-6Galβ1	27593.5	157
2,3Galβ1-4GlcNAcβ	27091.45	81, 82, 97, 141, 142
Galβ1-4GlcNAcβ	26658.48	152, 153
2Galβ1-4GlcNAcβ1	26061.04	60, 70
GlcNAcβ1-3Galβ1	24471.77	163, 164, 165, 166, 167
Fucα1-4GlcNAcβ	-24945.3	77
Galα1-4GlcNAcβ	-25069.7	112
(6OSO3)Galβ1-4GlcNAcβ	-25093.5	44
KDNa2-3Galβ1	-25176.4	187, 188

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant di-saccharide in CFG glycan array V2.0 are listed. The regression coefficients are obtained by PLS regression using di-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.0.

For VVL, we identified six di-saccharide sub-trees with significant positive regression coefficients (Table 5.4), which included all seven glycan chains with high binding intensities. All six di-saccharides involved terminal beta-linked GalNAcβ or

terminal alpha-linked GalNAcα. Previously, terminal beta-linked GalNAcβ was ranked high by both motif and intensity segregation methods, and terminal alpha-linked GalNAcα was only ranked high by the intensity segregation method [69]. In the CFG glycan array version 2.0, there are 10 glycan chains that have terminal alpha-linked GalNAcα and only two have high binding intensity with VVL. Moreover, there are 13 glycan chains that have terminal beta-linked GalNAcβ in the CFG glycan array version 2.0, and five of them have high binding intensity with VVL. Thus, using only terminal alpha- and beta- GalNAc to determine the binding specificity of VVL may be insufficient. When the terminal alpha- and beta- GalNAc are not attached directly to a spacer in the glycan array, our QSAR results implied that the saccharides attaching to terminal alpha- and beta-linked GalNAc affect the binding specificity to VVL. For example, in number 86 (GalNAcα1-3Galβ–Sp8) glycan chain of the CFG glycan array version 2.0, a terminal GalNAcα attached to a Gal with an α1-3 link and then attached to the spacer. This glycan chain has high binding intensity. However, when the galactose is also attached with a fucose, the glycan chain (number 84 (GalNAcα1-3(Fucα1-2)Galβ–Sp8) of the CFG glycan array version 2.0) loses its binding specificity. The QSAR method even showed that the GalNAcα1-2,3Galβ has a significant negative coefficient (Table 5.4). Similarly, a terminal GalNAc attached to a Gal with β1-3 linkage leads to high binding intensity as in number 89 (GalNAcβ1-3(Fucα1-2)Galβ-Sp8) and number 90 (GalNAcβ1-3Galα1-4Galβ1-4GlcNAcβ-Sp0) glycan chains of the CFG glycan array version 2.0. However, a terminal GalNAc attached to Gal with β1-4 linkage does not lead to high binding intensity as in five glycan chains of the CFG glycan array version 2.0:

- 203 (NeuAca2-8NeuAca2-8NeuAca2-8NeuAca2-3(GalNAcb1-4)Galb1-4Glcb-Sp0)
- 204 (Neu5Aca2-8Neu5Aca2-8Neu5Aca2-3(GalNAcb1-4)Galb1-4Glcb-Sp0)
- 206 (Neu5Aca2-8Neu5Acα2-3(GalNAcβ1-4)Galβ1-4Glcβ–Sp0)
- 209 (Neu5Aca2-3(GalNAcb1-4)Galb1-4GlcNAcb-Sp0)
- 210 (Neu5Aca2-3(GalNAcb1-4)Galb1-4GlcNAcb-Sp8)

Our results suggested that the binding specificity of VVL needs to be determined more carefully by considering the saccharides attached to the terminal GalNAc and how they are attached. This is consistent with the variance explained and R^2 results of the PLS regression study as the PLS regression using di-saccharide sub-trees as features gets much higher performance. Our results also suggested that visual inspection may not be able to identify the true binding specificities even for simple glycan array data. Furthermore, as shown in Table 5.4, each glycan chain contains only one significant di-saccharide, which suggests that all significant di-saccharides binding to VVL are independent.

For WGA, we identified 16 di-saccharide sub-trees with significant positive regression coefficients and four di-saccharide sub-trees with significant negative regression coefficient (Table 5.5). The 16 significant di-saccharides exist in 28 glycan chains of the CFG glycan array version 2.0, which all have high binding intensity with WGA. All 16 significant di-saccharides of WGA are independent as each glycan chain contains only one significant di-saccharide. There are three kinds of glycan in those 16 di-saccharides: Gal, GlcNAc and GalNAc. Those 16 di-saccharides cover the "terminal Lactosamine", "internal Lactosamine", "terminal GlcNAcβ", "Branching", and "terminal GalNAcα" motifs identified by motif and intensity segregation methods [69]. However, one highly ranked motif, "Terminal Neu5Acα2-3Gal", identified by motif and intensity segregation methods is missing. We then carefully examined the glycan array data of WGA. There are 37 glycan chains in CFG glycan array version 2.0 containing the "Terminal Neu5Acα2-3Gal" motif. The binding intensities of those 37 glycan chains to WGA are very broad, from -73 to 50264. It is likely that the di-saccharide sub-tree, "Terminal Neu5Acα2-3Gal", may not discern the binding specificity of those 37 glycan chains completely. We then used the tri- and tetra-saccharide sub-trees as features for the PLS regression. We were able to find one significant tri-saccharide (Neu5Acα2-3Galβ1-3GlcNAcβ) and four significant tetra-saccharide (Neu5Acα2-3Galβ1-3GlcNAcβ-Sp8, Neu5Acα2-3Galβ1-4(6OSO3)GlcNAcβ-Sp8, Neu5Acα2-3Galβ1-4GlcNAcβ-Sp0, Neu5Acα2-3Galβ1-4GlcNAcβ-Sp8) that contain terminal Neu5Acα2-3Gal. The results

implied that the terminal Neu5Acα2-3Gal may need to attach to a GlcNAc to make the binding to WGA more specific.

5.2.4 Evaluation of QSAR Model on Other Glycan-binding Proteins

To further demonstrate the effectiveness of the QSAR method, we tested it on several glycan-binding proteins. We first apply the QSAR to glycan array data of two plant lectins: Peanut Agglutinin (PNA) and Sambucus Nigra Lectin (SNA). For PNA, the PLS regressions using di-saccharide and tri-saccharide sub-trees as features can obtain significant high R^2 of 0.9603 and 0.9966 respectively (Table 5.1). We used PLS-Beta method to identify 11 tri-saccharide sub-trees with significant positive regression coefficients and three tri-saccharide sub-trees with significant negative regression coefficients (Table 5.6). Our results showed that PNA have high binding affinities with the terminal Galβ1–3GalNAc, which is consistent with previous study [73]. However, if there are other glycan attached to N-acetylgalactosamine (GalNAc) or the Galβ1–3GalNAc is directly attached to spacer arm (SP), the binding intensity of Galβ1–3GalNAc is reduced. We also observed that terminal galactose attached to another galactose with a β1-3 link also have high binding intensity. Furthermore, non-terminal Galβ1–3GalNAc has low binding affinity and leads to the negative coefficients. For SNA, the PLS regressions using di-saccharide sub-trees as features can obtain significant high R^2 of 0.9085 Table 5.1). As shown in Table 5.7, the PLS-Beta method successfully identified that SNA have high binding affinities with α2–6-linked *N*-Glycolylneuraminic Acid (Neu5Gcα2-6) bound to galactose (Gal) or N-acetylgalactosamine (GalNAc) [74].

Table 5.6 The significant tri-saccharide sub-trees binding specifically to PNA.

tri-saccharide	Regression Coefficient	Glycan numbers
(Galβ1-3)-3GalNAcα-Sp8	50301.69	125
(Galβ1-3)-3GalNAcβ-Sp8	49878.81	126
(Galβ1-3)-3Galβ-Sp8	47607.33	130
(Galβ1-3)-3,6GlcNAcα-Sp8	29692.31	120, 121, 122, 123
(Galβ1-3)-(3GalNAcβ1-4)-3,4Galβ1	28104.38	128, 201
(Galβ1-3)-3,6GalNAcα-Sp8	27781.27	150, 174, 241, 256
(Galβ1-3)-(3GalNAcβ1-3)-3Galα1	23494.39	127
(3GalNAcβ1-4)-(4Galβ1-4)-4Glcβ	15204.52	129
(4Galβ1-4)-4Glcβ-Sp8	15204.52	129
(Galβ1-3)-(3GalNAcβ1-4)-4Galβ1	15204.52	129
(Galβ1-3)(Neu5Acα-6)-3,6GalNAcα	13740.78	241
(3Galβ1-3)-(3GalNAcβ1-3)-3Galα1	-12988.4	223
(2Galβ1-3)-(3GalNAcβ1-4)-3,4Galβ1	-13998.6	59, 60
(Neu5Acα2-3)-(3Galβ1-3)-3GalNAcβ1	-19045.3	212, 223

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant tri-saccharide in CFG glycan array V2.0 are listed. The regression coefficients are obtained by PLS regression using tri-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.0.

Table 5.7 The significant di-saccharide sub-trees binding specifically to SNA.

di-saccharide	Regression Coefficient	Glycan numbers
(Neu5Gcα2-6)-6Galβ1	33419.7	263
(9NAcNeu5Acα2-6)-6Galβ1	29606.3	49
(Neu5Acα2-6)-6Galβ1	27618.6	53, 54, 244, 245, 246, 247, 248, 249, 250
(Neu5Acα2-6)-6Galβ	21657.7	251
(Neu5Acα2-6)-6GalNAcβ1	19823.3	243
(6GalNAcβ1-4)-4GlcNAcβ	19823.3	243

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant di-saccharide in CFG glycan array V2.0 are listed. The regression coefficients are obtained by PLS regression using di-

saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.0.

Then, we tested the QSAR method on the glycan array data of two animal lectins: sialic acid binding immunoglobulin-like lectin 8 (Siglec-8) and dendritic cell specific ICAM-3 grabbing non-integrin (DC-SIGN). For DC-SIGN, both PLS regression using di-saccharide and tri-saccharide sub-trees as features obtained significant high R^2 of 0.9179 and 0.9533 respectively (Table 5.1). The PLS-Beta method identified 22 tri-saccharide sub-trees with significant regression coefficients (Table 5.8). The top tri-saccharide, (Fucα1-3)(2Galβ1-4)-3,4GlcNAcβ (Le^X) is known to bind with DC-SIGN [75]. However, there is no mannose in our identified tri-saccharides. This may be because glycans containing mannose have both high and low binding intensities. For example, glycan chain 197 (Manα1-6(Manα1-3)Manα1-6(Manα1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn) and glycan chain 198 (Manα1-6(Manα1-3)Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn) in CFG glycan array V2.1 have similar structure. The glycan chain 197 is decomposed into 9 tri-saccharide sub-trees and glycan chain 198 is decomposed into 8 tri-saccharide sub-trees. Among them, six tri-saccharide sub-trees are the same. However, glycan chain 197 has a binding intensity of 6358.993 with DG-SIGN and glycan chain 198 have a binding intensity of 186.303 with DG-SIGN. For Siglec-8, PLS regression using all four types of sub-trees as features can obtain significant high R^2>0.99 (Table 5.1). The PLS-Beta method successfully identified the sialylated-sulfated galactose as the specific binding motif of Siglec-8 (Table 5.9).

Table 5.8 The significant tri-saccharide sub-trees binding specifically to DC-SIGN.

tri-saccharide	Regression Coefficient	Glycan numbers
(Fucα1-3)(2Galβ1-4)-3,4GlcNAcβ	4918.172	67, 68
(Fucα1-2)-(2Galβ1-4)-3,4GlcNAcβ	4918.172	67, 68
((3OSO3)Galβ1-4)-4(6OSO3)Glcβ-Sp0	4891.928	29
(GalNAcα1-3)-3GalNAcβ-Sp8	4706.763	85
(GlcNAcβ1-3)-(3Galβ1-4)-4GlcNAcβ1	4336.227	166
(3,4GalNAcβ1-3)-(3Galβ1-4)-3,4GlcNAcβ	4287.536	65
(2Galβ1-4)-3,4GlcNAcβ-Sp0	4044.822	67
((3OSO3)Galβ1-4)-4Glcβ-Sp8	4011.979	28
(3,4GlcNAcβ1-4)-(4Galβ1-4)-3,4GlcNAcβ1	3993.198	138
(4Galβ1-4)-(3,4GlcNAcβ1-4)-4Galβ1	3993.198	138
(Fucα1-3)(4Galβ1-4)-3,4GlcNAcβ1	3993.198	138
(Galβ1-4)-3,4GlcNAcβ-Sp8	3947.659	136
(KDNa2-3)-(3Galβ1-4)-4GlcNAcβ	3605.753	188
(4Galβ1-4)-4GlcNAcβ-Sp8	3303.441	170
(GlcNAcβ1-4)-(4Galβ1-4)-4GlcNAcβ	3303.441	170
(2Galβ1-4)-(3,4GalNAcβ1-3)-3Galβ1	3135.134	65, 66
(Fucα1-2)-(2Galβ1-4)-3,4GalNAcβ1	3135.134	65, 66
(Fucα1-3)(2Galβ1-4)-3,4GalNAcβ1	3135.134	65, 66
(Fucα1-3)-(3,4GalNAcβ1-3)-3Galβ1	3135.134	65, 66
(Fucα1-4)-3,4GlcNAc-Sp8	3036.607	118
(Galβ1-3)-3,4GlcNAc-Sp8	3036.607	118
(Fucα1-2)-2Galβ-Sp8	2947.905	74

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant tri-saccharide in CFG glycan array V2.1 are listed. The regression coefficients are obtained by PLS regression using tri-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V2.1.

Table 5.9 The significant mono-saccharide sub-trees binding specifically to Siglec-8.

mono-saccharide	Regression Coefficient	Glycan numbers
3(6OSO3)Galβ1	8655.456	246
3(6-O-Su)Galβ1	7816.967	354

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant mono-saccharide in CFG glycan array V4.1 are listed. The regression coefficients are obtained by PLS regression using mono-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V4.1.

Finally, we tested the QSAR method on the glycan array data of two antibodies: CSLEX1 (human CD15s antibody) and Sialyl Lewis x antibody (CD15s) - 10. For both antibodies, the PLS regression using tetra-saccharide sub-trees as features obtained the highest R^2>0.99 (Table 5.10). For CSLEX1, the PLS-Beta method successfully identified the Sialy1 Lewis x, ((Neu5Acα2-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ1, as the most significant tetra-saccharide sub-tree (Table 5.11). Meanwhile, the results showed that the glycan chain of (Neu5Acα2-3)-(3(6OSO3)Galβ1-4)-4GlcNAcβ also have binding affinity to the CSLEX1. For Sialyl Lewis x antibody (CD15s) - 10, the Sialy1 Lewis x, ((Neu5Acα2-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ1, is also among the top significant tetra-saccharide sub-trees (Table 5.12). It is not the most significant sub-tree because two glycan chains (number 297 (Galβ1-3(Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-6)GalNAcα-Sp14) and 377 (Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3GalNAcα-Sp14)) containing the Sialyl Lewis x structure have very low binding intensities. As shown in Table 5.12, the PLS regression assigned two other tetra-saccharide sub-trees [(3Galβ1-4)(Fucα1-3)-(3,4GlcNAcβ1-3)-3GalNAcα, (3Galβ1-4)-(3,4GlcNAcβ1-3)-3GalNAcα-Sp14] in glycan chain 377 with high negative coefficients in order to balance the high regression coefficient of Sialyl Lewis x. These results implied that the binding intensities of a Sialyl Lewis x are reduced if it is attached to an N-acetylgalactosamine.

Table 5.10 The R^2 of PLS regressions on glycan array data of different antibodies using different features.

	Mono	**Di**	**Tri**	**Tetra**
CSLEX1 (human CD15s antibody)	0.3954	0.6362	0.983	**0.9952**
Sialyl Lewis x antibody-10	0.1374	0.5147	0.9463	**0.9949**

Note: The highest values of R squares are highlighted in bold.

Table 5.11 The significant tetra-saccharide sub-trees binding specifically to CSLEX1 (human CD15s antibody).

tetra-saccharide	**Regression Coefficient**	**Glycan numbers**
((Neu5Acα2-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ1	11142.85	242, 245, 246, 290, 332, 374
(Neu5Acα2-3)-(3(6OSO3)Galβ1-4)-4GlcNAcβ-Sp8	9937.952	43
((Neu5Acα2-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ	9304.732	243, 244
(3Galb1-4)-(3,4GlcNAcβ1-3)-(3Galβ1-4)-3,4GlcNAcβ1	8146.454	242
((Neu5Acα2-3)-3(6-O-Su)Galb1-4)(Fucα1-3)-3,4GlcNAcβ	8145.086	220
(Fucα1-3)(3(6-O-Su)Galβ1-4)-3,4GlcNAcβ-Sp8	8145.086	220
(Neu5Acα2-3)-(3(6-O-Su)Galβ1-4)-3,4GlcNAcβ-Sp8	8145.086	220

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant tetra-saccharide in CFG glycan array V4.0 are listed. The regression coefficients are obtained by PLS regression using tetra-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V4.0.

Table 5.12 The significant tetra-saccharide sub-trees binding specifically to Sialyl Lewis x antibody-10.

tetra-saccharide	Regression Coefficient	Glycan numbers
(3Galβ1-4)-(3,4GlcNAcβ1-3)-(3Galβ1-4)-3,4GlcNAcβ1	1551.394	251
((Neu5Acα2-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ1	1424.837	251, 254, 255, 297, 377
(3Galβ1-4)(Fucα1-3)-(3,4GlcNAcβ1-3)-3Galβ	863.0468	254
(3Galβ1-4)-(3,4GlcNAcβ1-3)-3Galβ-Sp8	863.0468	254
(Fucα1-3)-(3,4GlcNAcβ1-3)-3Galβ-Sp8	863.0468	254
((3,4GlcNAcβ1-3)-3Galβ1-4)(Fucα1-3)-3,4GlcNAcβ1	654.8089	70, 251
(Fucα1-3)-(3,4GlcNAcβ1-3)-(3Galβ1-4)-3,4GlcNAcβ1	654.8089	70, 251
(3Galβ1-4)(Fucα1-3)-(3,4GlcNAcβ1-3)-3GalNAcα	-581.447	377
(3Galβ1-4)-(3,4GlcNAcβ1-3)-3GalNAcα-Sp14	-581.447	377
(2Galβ1-4)(Fucα1-3)-(3,4GlcNAcβ1-3)-3Galβ1	-702.1	69, 70
(2Galβ1-4)-(3,4GlcNAcβ1-3)-(3Galβ1-4)-3,4GlcNAcβ1	-896.585	70

Note: The sub-trees are ordered in descending of coefficients from up to bottom. The numbers of glycan chains including the significant tetra-saccharide in CFG glycan array V4.1 are listed. The regression coefficients are obtained by PLS regression using tetra-saccharide sub trees as features. The glycan numbers are the numbers of glycan chains in CFG glycan array V4.1.

5.3 Discussion

The application of glycan array is impeded currently by the lack of automatic and systematic methods to extract useful information [69]. In this study, we proposed a novel quantitative structure-activity relationship (QSAR) method to address this need. We first automatically decomposed the glycan chains into sub-trees. Then, we applied PLS

regression to the glycan array data using sub-trees as features. Based on PLS regression, we were able to identify significant sub-tees that contribute to binding. We demonstrated our methods on glycan array data of multiple glycan binding proteins. Moreover, the sub-structure features are generated automatically. We also developed a user-friendly web tool that can facilitate the rapid and automatic analysis of glycan array data.

Compared with predefined motifs, automatic decomposition of glycan chains into sub structures provides much broader features for selecting binding specificity. For example, in glycan array data of VVL, terminal alpha-linked GalNAc exists in glycan chains with both high and low binding intensities. Simply using terminal alpha-linked GalNAc as a feature to determine the binding specificity is insufficient. Actually, the motif segregation methods did not rank terminal alpha-linked GalNAc high. Meanwhile, by using di-saccharide sub-trees as features, our QSAR method successfully identified binding specific di-saccharides that include terminal alpha-linked GalNAc. Our results implied that the saccharide attached to terminal alpha-linked GalNAc also determined the binding specificity of VVL. Furthermore, the bindings of glycan chains to proteins are complex. Fixed features, like predefined motifs, may not be able to identify real binding specificity. For example, the QSAR method identified that a di-saccharide: 3,6Manβ1-4GlcNAcβ1, contributed to binding of ConA. This di-saccharide is a subset of both "N-glycan high mannose" and "N-Glycan complex" motifs. Further analysis showed that the 3,6Manβ1-4GlcNAcβ1 has a lower p-value based on motif segregation. Thus, the 3,6Manβ1-4GlcNAcβ1 may be the real contributor to the binding specificity of ConA.

Further experiments are needed to confirm the conclusion. But the QSAR method showed the potential to determine more representative binding specificities.

Both motif and intensity segregation methods need to separate the glycan data into two groups and then select the significant motif based on statistical tests on the intensities of the two groups. For intensity segregation, a threshold is needed to determine high and low intensities, which will bring uncertainty to the results [69]. Meanwhile, as the number of glycan chains containing a certain motif is low, the motif segregation suffers from unbalanced data in the two groups [69]. Our QSAR method overcomes the uncertainty and bias as it does not need to separate the glycan data into two groups as motif and intensity segregation methods.

Currently, we performed the PLS regression using different size sub-trees separately. Our current approach fixed the size of sub-structures to represent binding specificity. However, glycan-binding proteins may bind to different size sub-trees in glycan chains. For example, we showed that some di-saccharides of ConA are correlated and may be merged to tri-saccharide sub-trees. In the future, we will further develop the QSAR methods using all sub-trees under a certain size as features. However, directly using all sub-trees will lead to overfitting as features overlap. We are currently exploring feature selection methods to remove overlapped features. Then, the PLS regression will be performed on selected features.

In conclusion, our QSAR method provides a new tool for efficient analysis of glycan array data. Our method is general and can be applied to different types of glycan

array of different glycan-binding proteins. Our method should prompt the utilization of glycan array and help understand the biology of glycan-binding proteins.

Chapter 6

Conclusion and Future Work

In this thesis, we present a data-intensive computing platform for bioinformatics applications using virtualization technologies and HPC infrastructures. The platform seamlessly integrates the web user interface (presentation tier), scientific workflow (logic tier) and computing infrastructure (data/computing tier).

This platform is demonstrated through two bioinformatics projects. At first project, we redesigned and deployed CMD website using the Xen-based virtualization solution. Now CMD is a centralized web portal in the cotton research community. To achieve high-performance and scalability for CMD web tools, we hosted the large amounts of protein databases and computational intensive applications of CMD on the Palmetto HPC of Clemson University. Biologists can easily utilize both bioinformatics applications and HPC resources through the CMD website even with little background in computer science. Since 2009, CMD has been accessed by users from 101 countries all over the world and 48 states in USA. Currently, CMD project hosts 18 projects with 17,448 publicly available SSRs, 312 mapped cotton RFLP sequences containing SSRs and 27 genetic maps for *Gossypium* genomics.

Second, a web tool, Glycan Array QSAR Tool, was developed on our bioinformatics platform to analyze glycan array data. The user interface of this tool was developed at the top of Drupal Content Management Systems (CMS) and the computational part was implemented using MATLAB Compiler Runtime (MCR)

module. This QSAR method is general and can be applied to different types of glycan array of different glycan-binding proteins, and it should prompt the utilization of glycan array and help understanding the biology of glycan-binding proteins.

In conclusion, our new bioinformatics computing platform enables the rapid deployment of data-intensive bioinformatics applications using an HPC and virtualization environment with a user-friendly web interface and bridges the gap between biological scientists and cyberinfrastructure.

Some bioinformatics applications require specific computing environments, which are not provided by public computing infrastructures. For example, applications need a specific resource management system which is different with current HPC resources (e.g. SGE, Hadoop scheduler); prerequisite libraries or compliers conflict with its version in the current Operation System; the I/O access pattern does not match the storage system of HPC (e.g. Hadoop HDFS needs the shared-nothing file system). To overcome all difficulties related to the deployments of bioinformatics application on computing environment, our current solution is to port all those bioinformatics applications on the private computing infrastructures (e.g. remote private Linux cluster, local web server). However, for the larger-scale applications, the public HPC (e.g. Clemson Palmetto HPC) is the only feasible infrastructure which could provide sufficient computing resource. In our future research, we plan to extend the current virtualization solution from server virtualization to computing infrastructure virtualization. We will construct a virtual HPC cluster at the top of the public computing infrastructures, and

finally deploy all bioinformatics applications on public HPC or Cloud platforms. This new solution can bring many advantages:

- Improve the scalability of bioinformatics applications.
- Increase the flexibility of deployment and reduce the complexity of bioinformatics platform.
- Improve the reliability of whole system.
- Enhance the security of whole system

Appendix A

Table A-I. CFG array v2.0 mono-saccharide Features Summary

Number of Feature	mono-saccharide
0	(3OSO3)(6OSO3)Galb1
1	(3OSO3)Galb
2	(3OSO3)Galb1
3	(4OSO3)(6OSO3)Galb1
4	(4OSO3)Galb1
5	(6H2PO3)Mana
6	(6OSO3)Galb1
7	(6OSO3)GlcNAcb
8	2,3Galb
9	2,3Galb1
10	2,4Galb1
11	2Galb
12	2Galb1
13	2GlcNAcb
14	2Mana1
15	3(6-O-Su)Galb1
16	3(6OSO3)GalNAca
17	3(6OSO3)Galb1
18	3(6OSO3)GlcNAcb
19	3,4(6OSO3)GlcNAcb
20	3,4,6GlcNAc
21	3,4GalNAcb
22	3,4GalNAcb1
23	3,4Galb1
24	3,4GlcNAc
25	3,4GlcNAcb
26	3,4GlcNAcb1
27	3,6GalNAca
28	3,6GalNAcb
29	3,6Galb1
30	3,6GlcNAca
31	3,6GlcNAcb1
32	3,6Mana
33	3,6Mana1
34	3,6Manb1
35	3GalNAca
36	3GalNAcb
37	3GalNAcb1
38	3Gala
39	3Gala1
40	3Galb
41	3Galb1
42	3GlcNAca

43	3GlcNAcb
44	3GlcNAcb1
45	3Mana
46	4(6OSO3)GlcNAcb
47	4(6OSO3)Glcb
48	4,6GalNAca
49	4GalNAca1
50	4GalNAcb
51	4GalNAcb1
52	4Galb
53	4Galb1
54	4GlcNAcb
55	4GlcNAcb1
56	4Glca
57	4Glcb
58	4MDPLys
59	6GalNAca
60	6GalNAcb1
61	6Galb
62	6Galb1
63	6Glca1
64	6Glcb
65	8Neu5Aca
66	8Neu5Aca2
67	9NAcNeu5Aca
68	9NAcNeu5Aca2
69	Asn
70	Fuca1
71	Fucb1
72	GalNAca1
73	GalNAcb1
74	Gala1
75	Galb1
76	GlcAa
77	GlcAb
78	GlcAb1
79	GlcNAca1
80	GlcNAcb1
81	Glca1
82	Glcb1
83	Gly
84	KDNa2
85	Man5
86	Mana1
87	Manb1
88	Neu5Aca
89	Neu5Aca2
90	Neu5Acb2
91	Neu5Gca

92	Neu5Gca2
93	Sp0
94	Sp10
95	Sp11
96	Sp8
97	Sp9
98	a-D-Gal
99	a-D-Glc
100	a-D-Man
101	a-GalNAc
102	a-L-Fuc
103	a-L-Rha
104	a-NeuAc
105	b-D-Gal
106	b-D-Glc
107	b-D-Man
108	b-GalNAc
109	b-GlcN(Gc)
110	b-GlcNAc
111	b-NeuAc

Table A-II. CFG array v2.0 di-saccharide Features Summary

Number of Feature	di-saccharide
0	((3OSO3)(6OSO3)Galb1-4)-4(6OSO3)GlcNAcb
1	((3OSO3)(6OSO3)Galb1-4)-4GlcNAcb
2	((3OSO3)Galb1-3)-3,4GlcNAcb
3	((3OSO3)Galb1-3)-3GalNAca
4	((3OSO3)Galb1-3)-3GlcNAcb
5	((3OSO3)Galb1-4)-3,4GlcNAcb
6	((3OSO3)Galb1-4)-4(6OSO3)GlcNAcb
7	((3OSO3)Galb1-4)-4(6OSO3)Glcb
8	((3OSO3)Galb1-4)-4GlcNAcb
9	((3OSO3)Galb1-4)-4Glcb
10	((4OSO3)(6OSO3)Galb1-4)-4GlcNAcb
11	((4OSO3)Galb1-4)-4GlcNAcb
12	((6OSO3)Galb1-4)-4(6OSO3)Glcb
13	((6OSO3)Galb1-4)-4GlcNAcb
14	((6OSO3)Galb1-4)-4Glcb
15	(2,3Galb1-3)-3GlcNAcb
16	(2,3Galb1-4)-3,4GlcNAcb
17	(2,3Galb1-4)-4GlcNAcb
18	(2,3Galb1-4)-4Glcb
19	(2,4Galb1-4)-4GalNAcb
20	(2,4Galb1-4)-4GlcNAcb
21	(2Galb1-3)-3,4GlcNAcb
22	(2Galb1-3)-3GalNAca
23	(2Galb1-3)-3GalNAcb
24	(2Galb1-3)-3GalNAcb1
25	(2Galb1-3)-3GlcNAca
26	(2Galb1-4)-3,4GalNAcb1
27	(2Galb1-4)-3,4GlcNAcb
28	(2Galb1-4)-4GlcNAcb
29	(2Galb1-4)-4GlcNAcb1
30	(2Galb1-4)-4Glcb
31	(2Mana1-2)-2Mana1
32	(2Mana1-3)-3,6Mana
33	(2Mana1-3)-3,6Mana1
34	(2Mana1-3)-3,6Manb1
35	(2Mana1-3)-3Mana
36	(2Mana1-6)-3,6Mana
37	(2Mana1-6)-3,6Mana1
38	(2Mana1-6)-3,6Manb1
39	(3(6-O-Su)Galb1-4)-3,4GlcNAcb
40	(3(6OSO3)Galb1-4)-4GlcNAcb
41	(3,4GalNAcb1-3)-3Galb1
42	(3,4Galb1-4)-4GlcNAcb
43	(3,4Galb1-4)-4Glcb
44	(3,4GlcNAcb1-3)-3Galb
45	(3,4GlcNAcb1-3)-3Galb1
46	(3,4GlcNAcb1-4)-4Galb1

47	(3,6Galb1-4)-4GlcNAcb
48	(3,6GlcNAcb1-4)-4Galb1
49	(3,6Mana1-6)-3,6Manb1
50	(3,6Manb1-4)-4GlcNAcb1
51	(3GalNAcb1-3)-3Gala
52	(3GalNAcb1-3)-3Gala1
53	(3GalNAcb1-3)-3Galb1
54	(3GalNAcb1-4)-3,4Galb1
55	(3GalNAcb1-4)-4Galb1
56	(3GalNAcb1-4)-4GlcNAcb
57	(3Gala1-4)-4Galb1
58	(3Galb1-3)-3(6OSO3)GalNAca
59	(3Galb1-3)-3(6OSO3)GlcNAcb
60	(3Galb1-3)-3,4GlcNAcb
61	(3Galb1-3)-3,4GlcNAcb1
62	(3Galb1-3)-3,6GalNAcb
63	(3Galb1-3)-3GalNAca
64	(3Galb1-3)-3GalNAcb1
65	(3Galb1-3)-3GlcNAcb
66	(3Galb1-3)-3GlcNAcb1
67	(3Galb1-4)-3,4(6OSO3)GlcNAcb
68	(3Galb1-4)-3,4GalNAcb
69	(3Galb1-4)-3,4GlcNAcb
70	(3Galb1-4)-3,4GlcNAcb1
71	(3Galb1-4)-4(6OSO3)GlcNAcb
72	(3Galb1-4)-4GlcNAcb
73	(3Galb1-4)-4GlcNAcb1
74	(3Galb1-4)-4Glcb
75	(3GlcNAcb1-3)-3Galb1
76	(3OSO3)Galb-Sp8
77	(4GalNAca1-3)-2,3Galb1
78	(4GalNAcb1-3)-2,3Galb1
79	(4Galb1-4)-3,4GlcNAcb
80	(4Galb1-4)-3,4GlcNAcb1
81	(4Galb1-4)-4GalNAcb
82	(4Galb1-4)-4Galb
83	(4Galb1-4)-4GlcNAcb
84	(4Galb1-4)-4Glca
85	(4Galb1-4)-4Glcb
86	(4GlcNAcb1-2)-2Mana1
87	(4GlcNAcb1-3)-3,6GalNAca
88	(4GlcNAcb1-3)-3GalNAca
89	(4GlcNAcb1-3)-3Galb1
90	(4GlcNAcb1-4)-4GlcNAcb
91	(4GlcNAcb1-6)-3,6GalNAca
92	(4GlcNAcb1-6)-3,6GlcNAca
93	(4GlcNAcb1-6)-6GalNAca
94	(6GalNAcb1-4)-4GlcNAcb
95	(6Galb1-4)-4(6OSO3)GlcNAcb

96	(6Galb1-4)-4GlcNAcb
97	(6Galb1-4)-4GlcNAcb1
98	(6Galb1-4)-4Glcb
99	(6Glca1-6)-6Glcb
100	(6H2PO3)Mana-Sp8
101	(6OSO3)GlcNAcb-Sp8
102	(8Neu5Aca2-3)-3,4Galb1
103	(8Neu5Aca2-3)-3Galb1
104	(8Neu5Aca2-8)-8Neu5Aca
105	(8Neu5Aca2-8)-8Neu5Aca2
106	(9NAcNeu5Aca2-6)-6Galb1
107	(Fuca1-2)-2,3Galb
108	(Fuca1-2)-2,3Galb1
109	(Fuca1-2)-2,4Galb1
110	(Fuca1-2)-2Galb
111	(Fuca1-2)-2Galb1
112	(Fuca1-2)-2GlcNAcb
113	(Fuca1-3)-3,4(6OSO3)GlcNAcb
114	(Fuca1-3)-3,4GalNAcb
115	(Fuca1-3)-3,4GalNAcb1
116	(Fuca1-3)-3,4GlcNAcb
117	(Fuca1-3)-3,4GlcNAcb1
118	(Fuca1-3)-3GlcNAcb
119	(Fuca1-4)-3,4GlcNAc
120	(Fuca1-4)-3,4GlcNAcb
121	(Fuca1-4)-3,4GlcNAcb1
122	(Fuca1-4)-4GlcNAcb
123	(Fucb1-3)-3GlcNAcb
124	(GalNAca1-3)-2,3Galb
125	(GalNAca1-3)-2,3Galb1
126	(GalNAca1-3)-3GalNAca
127	(GalNAca1-3)-3GalNAcb
128	(GalNAca1-3)-3Galb
129	(GalNAca1-4)-2,4Galb1
130	(GalNAcb1-3)-2,3Galb
131	(GalNAcb1-3)-3Gala1
132	(GalNAcb1-4)-3,4Galb1
133	(GalNAcb1-4)-3,4GlcNAcb
134	(GalNAcb1-4)-4GlcNAcb
135	(Gala1-2)-2Galb
136	(Gala1-3)-2,3Galb
137	(Gala1-3)-2,3Galb1
138	(Gala1-3)-3,4Galb1
139	(Gala1-3)-3Galb
140	(Gala1-3)-3Galb1
141	(Gala1-3)-3GlcNAca
142	(Gala1-3)-3GlcNAcb
143	(Gala1-4)-2,4Galb1
144	(Gala1-4)-3,4Galb1

145	(Gala1-4)-4Galb1
146	(Gala1-4)-4GlcNAcb
147	(Gala1-6)-6Glcb
148	(Galb1-2)-2Galb
149	(Galb1-3)-3,4GlcNAc
150	(Galb1-3)-3,4GlcNAcb
151	(Galb1-3)-3,4GlcNAcb1
152	(Galb1-3)-3,6GalNAca
153	(Galb1-3)-3,6GlcNAca
154	(Galb1-3)-3,6GlcNAcb1
155	(Galb1-3)-3GalNAca
156	(Galb1-3)-3GalNAcb
157	(Galb1-3)-3GalNAcb1
158	(Galb1-3)-3Galb
159	(Galb1-3)-3GlcNAcb
160	(Galb1-4)-3,4GlcNAcb
161	(Galb1-4)-3,4GlcNAcb1
162	(Galb1-4)-4(6OSO3)Glcb
163	(Galb1-4)-4GalNAca1
164	(Galb1-4)-4GalNAcb1
165	(Galb1-4)-4GlcNAcb
166	(Galb1-4)-4GlcNAcb1
167	(Galb1-4)-4Glcb
168	(GlcAb1-3)-3Galb
169	(GlcAb1-6)-6Galb
170	(GlcNAca1-3)-3Galb1
171	(GlcNAca1-6)-6Galb1
172	(GlcNAcb1-2)-2Galb1
173	(GlcNAcb1-2)-2Mana1
174	(GlcNAcb1-3)-3,4,6GlcNAc
175	(GlcNAcb1-3)-3,6Galb1
176	(GlcNAcb1-3)-3,6GlcNAca
177	(GlcNAcb1-3)-3GalNAca
178	(GlcNAcb1-3)-3Galb
179	(GlcNAcb1-3)-3Galb1
180	(GlcNAcb1-4)-3,4,6GlcNAc
181	(GlcNAcb1-4)-4,6GalNAca
182	(GlcNAcb1-4)-4Galb1
183	(GlcNAcb1-4)-4GlcNAcb1
184	(GlcNAcb1-6)-3,4,6GlcNAc
185	(GlcNAcb1-6)-3,6GalNAca
186	(GlcNAcb1-6)-3,6Galb1
187	(GlcNAcb1-6)-3,6GlcNAca
188	(GlcNAcb1-6)-4,6GalNAca
189	(GlcNAcb1-6)-6GalNAca
190	(GlcNAcb1-6)-6Galb1
191	(Glca1-4)-4Glca
192	(Glca1-4)-4Glcb
193	(Glca1-6)-6Glca1

194	(Glcb1-4)-4Glcb
195	(Glcb1-6)-6Glcb
196	(KDNa2-3)-3Galb1
197	(Mana1-2)-2Mana1
198	(Mana1-3)-3,6Mana
199	(Mana1-3)-3,6Mana1
200	(Mana1-3)-3,6Manb1
201	(Mana1-6)-3,6Mana
202	(Mana1-6)-3,6Mana1
203	(Mana1-6)-3,6Manb1
204	(Manb1-4)-4GlcNAcb
205	(Neu5Aca2-3)-3(6-O-Su)Galb1
206	(Neu5Aca2-3)-3(6OSO3)Galb1
207	(Neu5Aca2-3)-3,4Galb1
208	(Neu5Aca2-3)-3,6GalNAca
209	(Neu5Aca2-3)-3GalNAca
210	(Neu5Aca2-3)-3GalNAcb1
211	(Neu5Aca2-3)-3Galb
212	(Neu5Aca2-3)-3Galb1
213	(Neu5Aca2-6)-3,6GalNAca
214	(Neu5Aca2-6)-3,6GalNAcb
215	(Neu5Aca2-6)-3,6GlcNAca
216	(Neu5Aca2-6)-3,6GlcNAcb1
217	(Neu5Aca2-6)-6GalNAca
218	(Neu5Aca2-6)-6GalNAcb1
219	(Neu5Aca2-6)-6Galb
220	(Neu5Aca2-6)-6Galb1
221	(Neu5Aca2-8)-8Neu5Aca
222	(Neu5Aca2-8)-8Neu5Aca2
223	(Neu5Acb2-6)-3,6GalNAca
224	(Neu5Acb2-6)-3,6GlcNAca
225	(Neu5Acb2-6)-6GalNAca
226	(Neu5Acb2-6)-6Galb1
227	(Neu5Gca2-3)-3Galb1
228	(Neu5Gca2-6)-6GalNAca
229	(Neu5Gca2-6)-6Galb1
230	2,3Galb-Sp8
231	2Galb-Sp8
232	2GlcNAcb-Sp8
233	3(6OSO3)GalNAca-Sp8
234	3(6OSO3)GlcNAcb-Sp8
235	3,4(6OSO3)GlcNAcb-Sp8
236	3,4,6GlcNAc-Sp8
237	3,4GalNAcb-Sp0
238	3,4GalNAcb-Sp8
239	3,4GlcNAc-Sp0
240	3,4GlcNAc-Sp8
241	3,4GlcNAcb-Sp0
242	3,4GlcNAcb-Sp8

243	3,6GalNAca-Sp8
244	3,6GalNAcb-Sp8
245	3,6GlcNAca-Sp8
246	3,6Mana-Sp9
247	3GalNAca-Sp8
248	3GalNAcb-Sp0
249	3GalNAcb-Sp8
250	3Gala-Sp9
251	3Galb-Sp8
252	3GlcNAca-Sp8
253	3GlcNAcb-Sp0
254	3GlcNAcb-Sp8
255	3Mana-Sp9
256	4(6OSO3)GlcNAcb-Sp0
257	4(6OSO3)GlcNAcb-Sp8
258	4(6OSO3)Glcb-Sp0
259	4(6OSO3)Glcb-Sp8
260	4,6GalNAca-Sp8
261	4GalNAcb-Sp0
262	4GalNAcb-Sp8
263	4Galb-Sp0
264	4GlcNAcb-Asn
265	4GlcNAcb-Gly
266	4GlcNAcb-Sp0
267	4GlcNAcb-Sp8
268	4GlcNAcb1-Sp8
269	4Glca-Sp8
270	4Glca-Sp9
271	4Glcb-Sp0
272	4Glcb-Sp10
273	4Glcb-Sp8
274	4Glcb-Sp9
275	6GalNAca-Sp0
276	6GalNAca-Sp8
277	6Galb-Sp8
278	6Glcb-Sp8
279	8Neu5Aca-Sp8
280	9NAcNeu5Aca-Sp8
281	GlcAa-Sp8
282	GlcAb-Sp8
283	GlcNAcb1-4MDPLys
284	Man5-Asn
285	Neu5Aca-Sp11
286	Neu5Gca-Sp8
287	a-D-Gal-Sp8
288	a-D-Glc-Sp8
289	a-D-Man-Sp8
290	a-GalNAc-Sp8
291	a-L-Fuc-Sp8

292	a-L-Fuc-Sp9
293	a-L-Rha-Sp8
294	a-NeuAc-Sp8
295	b-D-Gal-Sp8
296	b-D-Glc-Sp8
297	b-D-Man-Sp8
298	b-GalNAc-Sp8
299	b-GlcN(Gc)-Sp8
300	b-GlcNAc-Sp0
301	b-GlcNAc-Sp8
302	b-NeuAc-Sp8

Table A-III. CFG array v2.0 tri-saccharide Features Summary

Number of Feature	tri-saccharide
0	((3OSO3)(6OSO3)Galb1-4)-4(6OSO3)GlcNAcb-Sp0
1	((3OSO3)(6OSO3)Galb1-4)-4GlcNAcb-Sp0
2	((3OSO3)Galb1-3)(Fuca1-4)-3,4GlcNAcb
3	((3OSO3)Galb1-3)-3,4GlcNAcb-Sp8
4	((3OSO3)Galb1-3)-3GalNAca-Sp8
5	((3OSO3)Galb1-3)-3GlcNAcb-Sp8
6	((3OSO3)Galb1-4)-3,4GlcNAcb-Sp8
7	((3OSO3)Galb1-4)-4(6OSO3)GlcNAcb-Sp8
8	((3OSO3)Galb1-4)-4(6OSO3)Glcb-Sp0
9	((3OSO3)Galb1-4)-4(6OSO3)Glcb-Sp8
10	((3OSO3)Galb1-4)-4GlcNAcb-Sp0
11	((3OSO3)Galb1-4)-4GlcNAcb-Sp8
12	((3OSO3)Galb1-4)-4Glcb-Sp8
13	((4OSO3)(6OSO3)Galb1-4)-4GlcNAcb-Sp0
14	((4OSO3)Galb1-4)-4GlcNAcb-Sp8
15	((6OSO3)Galb1-4)-4(6OSO3)Glcb-Sp8
16	((6OSO3)Galb1-4)-4GlcNAcb-Sp8
17	((6OSO3)Galb1-4)-4Glcb-Sp0
18	((6OSO3)Galb1-4)-4Glcb-Sp8
19	(2,3Galb1-3)-3GlcNAcb-Sp0
20	(2,3Galb1-4)-3,4GlcNAcb-Sp0
21	(2,3Galb1-4)-4GlcNAcb-Sp0
22	(2,3Galb1-4)-4GlcNAcb-Sp8
23	(2,3Galb1-4)-4Glcb-Sp0
24	(2,4Galb1-4)-4GalNAcb-Sp8
25	(2,4Galb1-4)-4GlcNAcb-Sp8
26	(2Galb1-3)(Fuca1-4)-3,4GlcNAcb
27	(2Galb1-3)-(3GalNAcb1-3)-3Gala
28	(2Galb1-3)-(3GalNAcb1-3)-3Gala1
29	(2Galb1-3)-(3GalNAcb1-3)-3Galb1
30	(2Galb1-3)-(3GalNAcb1-4)-3,4Galb1
31	(2Galb1-3)-3,4GlcNAcb-Sp8
32	(2Galb1-3)-3GalNAca-Sp8
33	(2Galb1-3)-3GalNAcb-Sp0
34	(2Galb1-3)-3GalNAcb-Sp8
35	(2Galb1-3)-3GlcNAca-Sp8
36	(2Galb1-4)-(3,4GalNAcb1-3)-3Galb1
37	(2Galb1-4)-(4GlcNAcb1-3)-3Galb1
38	(2Galb1-4)-3,4GlcNAcb-Sp0
39	(2Galb1-4)-3,4GlcNAcb-Sp8
40	(2Galb1-4)-4GlcNAcb-Sp0
41	(2Galb1-4)-4GlcNAcb-Sp8
42	(2Galb1-4)-4Glcb-Sp0
43	(2Mana1-2)-(2Mana1-3)-3,6Manb1
44	(2Mana1-2)-(2Mana1-3)-3Mana
45	(2Mana1-2)-(2Mana1-6)-3,6Mana
46	(2Mana1-3)(2Mana1-6)-3,6Mana

47	(2Mana1-3)(2Mana1-6)-3,6Mana1
48	(2Mana1-3)(2Mana1-6)-3,6Manb1
49	(2Mana1-3)(3,6Mana1-6)-3,6Manb1
50	(2Mana1-3)(Mana1-6)-3,6Mana1
51	(2Mana1-3)-(3,6Mana1-6)-3,6Manb1
52	(2Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
53	(2Mana1-3)-3,6Mana-Sp9
54	(2Mana1-3)-3Mana-Sp9
55	(2Mana1-6)-(3,6Mana1-6)-3,6Manb1
56	(2Mana1-6)-(3,6Manb1-4)-4GlcNAcb1
57	(2Mana1-6)-3,6Mana-Sp9
58	(3(6-O-Su)Galb1-4)-3,4GlcNAcb-Sp8
59	(3(6OSO3)Galb1-4)-4GlcNAcb-Sp8
60	(3,4GalNAcb1-3)-(3Galb1-4)-3,4GlcNAcb
61	(3,4GalNAcb1-3)-(3Galb1-4)-3,4GlcNAcb1
62	(3,4Galb1-4)-4GlcNAcb-Sp0
63	(3,4Galb1-4)-4GlcNAcb-Sp8
64	(3,4Galb1-4)-4Glcb-Sp0
65	(3,4Galb1-4)-4Glcb-Sp9
66	(3,4GlcNAcb1-3)-(3Galb1-4)-3,4GalNAcb
67	(3,4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb
68	(3,4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb1
69	(3,4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb
70	(3,4GlcNAcb1-3)-3Galb-Sp8
71	(3,4GlcNAcb1-4)-(4Galb1-4)-3,4GlcNAcb
72	(3,4GlcNAcb1-4)-(4Galb1-4)-3,4GlcNAcb1
73	(3,6Galb1-4)-4GlcNAcb-Sp8
74	(3,6GlcNAcb1-4)-(4Galb1-4)-4Glcb
75	(3,6Mana1-6)-(3,6Manb1-4)-4GlcNAcb1
76	(3,6Manb1-4)-(4GlcNAcb1-4)-4GlcNAcb
77	(3GalNAcb1-3)-(3Gala1-4)-4Galb1
78	(3GalNAcb1-3)-(3Galb1-4)-4GlcNAcb
79	(3GalNAcb1-3)-(3Galb1-4)-4Glcb
80	(3GalNAcb1-3)-3Gala-Sp9
81	(3GalNAcb1-4)-(3,4Galb1-4)-4Glcb
82	(3GalNAcb1-4)-(4Galb1-4)-4Glcb
83	(3GalNAcb1-4)-4GlcNAcb-Sp0
84	(3Gala1-4)-(4Galb1-4)-4GlcNAcb
85	(3Gala1-4)-(4Galb1-4)-4Glca
86	(3Gala1-4)-(4Galb1-4)-4Glcb
87	(3Galb1-3)(3Galb1-4)-3,4GlcNAcb
88	(3Galb1-3)(Fuca1-4)-3,4GlcNAcb
89	(3Galb1-3)(Fuca1-4)-3,4GlcNAcb1
90	(3Galb1-3)(Neu5Aca2-6)-3,6GalNAcb
91	(3Galb1-3)-(3,4GlcNAcb1-3)-3Galb1
92	(3Galb1-3)-(3GalNAcb1-3)-3Gala1
93	(3Galb1-3)-(3GalNAcb1-4)-3,4Galb1
94	(3Galb1-3)-(3GlcNAcb1-3)-3Galb1
95	(3Galb1-3)-3(6OSO3)GalNAca-Sp8

96	(3Galb1-3)-3(6OSO3)GlcNAcb-Sp8
97	(3Galb1-3)-3,4GlcNAcb-Sp0
98	(3Galb1-3)-3,4GlcNAcb-Sp8
99	(3Galb1-3)-3,6GalNAcb-Sp8
100	(3Galb1-3)-3GalNAca-Sp8
101	(3Galb1-3)-3GlcNAcb-Sp0
102	(3Galb1-3)-3GlcNAcb-Sp8
103	(3Galb1-4)-(3,4GlcNAcb1-3)-3Galb
104	(3Galb1-4)-(3,4GlcNAcb1-3)-3Galb1
105	(3Galb1-4)-(4GlcNAcb1-3)-3Galb1
106	(3Galb1-4)-3,4(6OSO3)GlcNAcb-Sp8
107	(3Galb1-4)-3,4GalNAcb-Sp0
108	(3Galb1-4)-3,4GalNAcb-Sp8
109	(3Galb1-4)-3,4GlcNAcb-Sp0
110	(3Galb1-4)-3,4GlcNAcb-Sp8
111	(3Galb1-4)-4(6OSO3)GlcNAcb-Sp8
112	(3Galb1-4)-4GlcNAcb-Sp0
113	(3Galb1-4)-4GlcNAcb-Sp8
114	(3Galb1-4)-4Glcb-Sp0
115	(3Galb1-4)-4Glcb-Sp10
116	(3Galb1-4)-4Glcb-Sp8
117	(3GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb
118	(4GalNAca1-3)-(2,3Galb1-4)-4GlcNAcb
119	(4GalNAcb1-3)-(2,3Galb1-4)-4GlcNAcb
120	(4Galb1-4)-(3,4GlcNAcb1-4)-4Galb1
121	(4Galb1-4)-3,4GlcNAcb-Sp0
122	(4Galb1-4)-4GalNAcb-Sp0
123	(4Galb1-4)-4GalNAcb-Sp8
124	(4Galb1-4)-4Galb-Sp0
125	(4Galb1-4)-4GlcNAcb-Sp0
126	(4Galb1-4)-4GlcNAcb-Sp8
127	(4Galb1-4)-4Glca-Sp9
128	(4Galb1-4)-4Glcb-Sp0
129	(4Galb1-4)-4Glcb-Sp10
130	(4Galb1-4)-4Glcb-Sp8
131	(4GlcNAcb1-2)-(2Mana1-3)-3,6Manb1
132	(4GlcNAcb1-2)-(2Mana1-6)-3,6Manb1
133	(4GlcNAcb1-3)(4GlcNAcb1-6)-3,6GalNAca
134	(4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb
135	(4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb1
136	(4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb
137	(4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb1
138	(4GlcNAcb1-3)-(3Galb1-4)-4Glcb
139	(4GlcNAcb1-3)-3,6GalNAca-Sp8
140	(4GlcNAcb1-3)-3GalNAca-Sp8
141	(4GlcNAcb1-4)-4GlcNAcb-Asn
142	(4GlcNAcb1-4)-4GlcNAcb-Gly
143	(4GlcNAcb1-4)-4GlcNAcb-Sp8
144	(4GlcNAcb1-6)-3,6GalNAca-Sp8

145	(4GlcNAcb1-6)-3,6GlcNAca-Sp8
146	(4GlcNAcb1-6)-6GalNAca-Sp8
147	(6GalNAcb1-4)-4GlcNAcb-Sp0
148	(6Galb1-4)-(4GlcNAcb1-2)-2Mana1
149	(6Galb1-4)-(4GlcNAcb1-3)-3Galb1
150	(6Galb1-4)-4(6OSO3)GlcNAcb-Sp8
151	(6Galb1-4)-4GlcNAcb-Sp0
152	(6Galb1-4)-4GlcNAcb-Sp8
153	(6Galb1-4)-4Glcb-Sp0
154	(6Galb1-4)-4Glcb-Sp8
155	(6Glca1-6)-6Glcb-Sp8
156	(8Neu5Aca2-3)(GalNAcb1-4)-3,4Galb1
157	(8Neu5Aca2-3)-(3,4Galb1-4)-4Glcb
158	(8Neu5Aca2-3)-(3Galb1-4)-4Glcb
159	(8Neu5Aca2-8)-(8Neu5Aca2-3)-3,4Galb1
160	(8Neu5Aca2-8)-(8Neu5Aca2-3)-3Galb1
161	(8Neu5Aca2-8)-8Neu5Aca-Sp8
162	(9NAcNeu5Aca2-6)-(6Galb1-4)-4GlcNAcb
163	(Fuca1-2)(4GalNAca1-3)-2,3Galb1
164	(Fuca1-2)(4GalNAcb1-3)-2,3Galb1
165	(Fuca1-2)(GalNAca1-3)-2,3Galb
166	(Fuca1-2)(GalNAca1-3)-2,3Galb1
167	(Fuca1-2)(GalNAca1-4)-2,4Galb1
168	(Fuca1-2)(GalNAcb1-3)-2,3Galb
169	(Fuca1-2)(Gala1-3)-2,3Galb
170	(Fuca1-2)(Gala1-3)-2,3Galb1
171	(Fuca1-2)(Gala1-4)-2,4Galb1
172	(Fuca1-2)-(2,3Galb1-3)-3GlcNAcb
173	(Fuca1-2)-(2,3Galb1-4)-3,4GlcNAcb
174	(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb
175	(Fuca1-2)-(2,3Galb1-4)-4Glcb
176	(Fuca1-2)-(2,4Galb1-4)-4GalNAcb
177	(Fuca1-2)-(2,4Galb1-4)-4GlcNAcb
178	(Fuca1-2)-(2Galb1-3)-3,4GlcNAcb
179	(Fuca1-2)-(2Galb1-3)-3GalNAca
180	(Fuca1-2)-(2Galb1-3)-3GalNAcb
181	(Fuca1-2)-(2Galb1-3)-3GalNAcb1
182	(Fuca1-2)-(2Galb1-4)-3,4GalNAcb1
183	(Fuca1-2)-(2Galb1-4)-3,4GlcNAcb
184	(Fuca1-2)-(2Galb1-4)-4GlcNAcb
185	(Fuca1-2)-(2Galb1-4)-4GlcNAcb1
186	(Fuca1-2)-(2Galb1-4)-4Glcb
187	(Fuca1-2)-2,3Galb-Sp8
188	(Fuca1-2)-2Galb-Sp8
189	(Fuca1-2)-2GlcNAcb-Sp8
190	(Fuca1-3)((3OSO3)Galb1-4)-3,4GlcNAcb
191	(Fuca1-3)(2,3Galb1-4)-3,4GlcNAcb
192	(Fuca1-3)(2Galb1-4)-3,4GalNAcb1
193	(Fuca1-3)(2Galb1-4)-3,4GlcNAcb

194	(Fuca1-3)(3(6-O-Su)Galb1-4)-3,4GlcNAcb
195	(Fuca1-3)(3Galb1-4)-3,4(6OSO3)GlcNAcb
196	(Fuca1-3)(3Galb1-4)-3,4GalNAcb
197	(Fuca1-3)(3Galb1-4)-3,4GlcNAcb
198	(Fuca1-3)(3Galb1-4)-3,4GlcNAcb1
199	(Fuca1-3)(4Galb1-4)-3,4GlcNAcb
200	(Fuca1-3)(4Galb1-4)-3,4GlcNAcb1
201	(Fuca1-3)(GalNAcb1-4)-3,4GlcNAcb
202	(Fuca1-3)(Galb1-4)-3,4GlcNAcb
203	(Fuca1-3)(Galb1-4)-3,4GlcNAcb1
204	(Fuca1-3)-(3,4GalNAcb1-3)-3Galb1
205	(Fuca1-3)-(3,4GlcNAcb1-3)-3Galb
206	(Fuca1-3)-(3,4GlcNAcb1-3)-3Galb1
207	(Fuca1-3)-(3,4GlcNAcb1-4)-4Galb1
208	(Fuca1-3)-3,4(6OSO3)GlcNAcb-Sp8
209	(Fuca1-3)-3,4GalNAcb-Sp0
210	(Fuca1-3)-3,4GalNAcb-Sp8
211	(Fuca1-3)-3,4GlcNAcb-Sp0
212	(Fuca1-3)-3,4GlcNAcb-Sp8
213	(Fuca1-3)-3GlcNAcb-Sp8
214	(Fuca1-4)-(3,4GlcNAcb1-3)-3Galb1
215	(Fuca1-4)-3,4GlcNAc-Sp0
216	(Fuca1-4)-3,4GlcNAc-Sp8
217	(Fuca1-4)-3,4GlcNAcb-Sp0
218	(Fuca1-4)-3,4GlcNAcb-Sp8
219	(Fuca1-4)-4GlcNAcb-Sp8
220	(Fucb1-3)-3GlcNAcb-Sp8
221	(GalNAca1-3)-(2,3Galb1-3)-3GlcNAcb
222	(GalNAca1-3)-(2,3Galb1-4)-3,4GlcNAcb
223	(GalNAca1-3)-(2,3Galb1-4)-4GlcNAcb
224	(GalNAca1-3)-(2,3Galb1-4)-4Glcb
225	(GalNAca1-3)-2,3Galb-Sp8
226	(GalNAca1-3)-3GalNAca-Sp8
227	(GalNAca1-3)-3GalNAcb-Sp8
228	(GalNAca1-3)-3Galb-Sp8
229	(GalNAca1-4)-(2,4Galb1-4)-4GlcNAcb
230	(GalNAcb1-3)-(3Gala1-4)-4Galb1
231	(GalNAcb1-3)-2,3Galb-Sp8
232	(GalNAcb1-4)-(3,4Galb1-4)-4GlcNAcb
233	(GalNAcb1-4)-(3,4Galb1-4)-4Glcb
234	(GalNAcb1-4)-3,4GlcNAcb-Sp0
235	(GalNAcb1-4)-4GlcNAcb-Sp0
236	(GalNAcb1-4)-4GlcNAcb-Sp8
237	(Gala1-2)-2Galb-Sp8
238	(Gala1-3)(Gala1-4)-3,4Galb1
239	(Gala1-3)-(2,3Galb1-3)-3GlcNAcb
240	(Gala1-3)-(2,3Galb1-4)-3,4GlcNAcb
241	(Gala1-3)-(2,3Galb1-4)-4GlcNAcb
242	(Gala1-3)-(2,3Galb1-4)-4Glcb

243	(Gala1-3)-(3,4Galb1-4)-4GlcNAcb
244	(Gala1-3)-(3Galb1-3)-3GlcNAcb
245	(Gala1-3)-(3Galb1-4)-3,4GalNAcb
246	(Gala1-3)-(3Galb1-4)-4GlcNAcb
247	(Gala1-3)-(3Galb1-4)-4Glcb
248	(Gala1-3)-2,3Galb-Sp8
249	(Gala1-3)-3Galb-Sp8
250	(Gala1-3)-3GlcNAca-Sp8
251	(Gala1-3)-3GlcNAcb-Sp8
252	(Gala1-4)-(2,4Galb1-4)-4GalNAcb
253	(Gala1-4)-(3,4Galb1-4)-4GlcNAcb
254	(Gala1-4)-(4Galb1-4)-4GalNAcb
255	(Gala1-4)-(4Galb1-4)-4Galb
256	(Gala1-4)-4GlcNAcb-Sp8
257	(Gala1-6)-6Glcb-Sp8
258	(Galb1-2)-2Galb-Sp8
259	(Galb1-3)(4GlcNAcb1-6)-3,6GalNAca
260	(Galb1-3)(4GlcNAcb1-6)-3,6GlcNAca
261	(Galb1-3)(Fuca1-4)-3,4GlcNAc
262	(Galb1-3)(Fuca1-4)-3,4GlcNAcb
263	(Galb1-3)(Fuca1-4)-3,4GlcNAcb1
264	(Galb1-3)(GlcNAcb1-6)-3,6GalNAca
265	(Galb1-3)(GlcNAcb1-6)-3,6GlcNAca
266	(Galb1-3)(Neu5Aca2-6)-3,6GalNAca
267	(Galb1-3)(Neu5Aca2-6)-3,6GlcNAca
268	(Galb1-3)(Neu5Aca2-6)-3,6GlcNAcb1
269	(Galb1-3)(Neu5Acb2-6)-3,6GalNAca
270	(Galb1-3)(Neu5Acb2-6)-3,6GlcNAca
271	(Galb1-3)-(3,4GlcNAcb1-3)-3Galb1
272	(Galb1-3)-(3,6GlcNAcb1-4)-4Galb1
273	(Galb1-3)-(3GalNAcb1-3)-3Gala1
274	(Galb1-3)-(3GalNAcb1-3)-3Galb1
275	(Galb1-3)-(3GalNAcb1-4)-3,4Galb1
276	(Galb1-3)-(3GalNAcb1-4)-4Galb1
277	(Galb1-3)-3,4GlcNAc-Sp0
278	(Galb1-3)-3,4GlcNAc-Sp8
279	(Galb1-3)-3,4GlcNAcb-Sp8
280	(Galb1-3)-3,6GalNAca-Sp8
281	(Galb1-3)-3,6GlcNAca-Sp8
282	(Galb1-3)-3GalNAca-Sp8
283	(Galb1-3)-3GalNAcb-Sp8
284	(Galb1-3)-3Galb-Sp8
285	(Galb1-3)-3GlcNAcb-Sp0
286	(Galb1-3)-3GlcNAcb-Sp8
287	(Galb1-4)-(3,4GlcNAcb1-4)-4Galb1
288	(Galb1-4)-(4GalNAca1-3)-2,3Galb1
289	(Galb1-4)-(4GalNAcb1-3)-2,3Galb1
290	(Galb1-4)-(4GlcNAcb1-2)-2Mana1
291	(Galb1-4)-(4GlcNAcb1-3)-3,6GalNAca

292	(Galb1-4)-(4GlcNAcb1-3)-3GalNAca
293	(Galb1-4)-(4GlcNAcb1-3)-3Galb1
294	(Galb1-4)-(4GlcNAcb1-6)-3,6GalNAca
295	(Galb1-4)-(4GlcNAcb1-6)-3,6GlcNAca
296	(Galb1-4)-(4GlcNAcb1-6)-6GalNAca
297	(Galb1-4)-3,4GlcNAcb-Sp0
298	(Galb1-4)-3,4GlcNAcb-Sp8
299	(Galb1-4)-4(6OSO3)Glcb-Sp0
300	(Galb1-4)-4(6OSO3)Glcb-Sp8
301	(Galb1-4)-4GlcNAcb-Sp0
302	(Galb1-4)-4GlcNAcb-Sp8
303	(Galb1-4)-4Glcb-Sp0
304	(Galb1-4)-4Glcb-Sp8
305	(GlcAb1-3)-3Galb-Sp8
306	(GlcAb1-6)-6Galb-Sp8
307	(GlcNAca1-3)-(3Galb1-4)-4GlcNAcb
308	(GlcNAca1-6)-(6Galb1-4)-4GlcNAcb
309	(GlcNAcb1-2)-(2Galb1-3)-3GlcNAca
310	(GlcNAcb1-2)-(2Mana1-3)-3,6Manb1
311	(GlcNAcb1-2)-(2Mana1-6)-3,6Manb1
312	(GlcNAcb1-3)(GlcNAcb1-4)-3,4,6GlcNAc
313	(GlcNAcb1-3)(GlcNAcb1-6)-3,4,6GlcNAc
314	(GlcNAcb1-3)(GlcNAcb1-6)-3,6Galb1
315	(GlcNAcb1-3)(GlcNAcb1-6)-3,6GlcNAca
316	(GlcNAcb1-3)-(3,6Galb1-4)-4GlcNAcb
317	(GlcNAcb1-3)-(3Galb1-3)-3GalNAca
318	(GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb
319	(GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb1
320	(GlcNAcb1-3)-(3Galb1-4)-4Glcb
321	(GlcNAcb1-3)-3,4,6GlcNAc-Sp8
322	(GlcNAcb1-3)-3,6GlcNAca-Sp8
323	(GlcNAcb1-3)-3GalNAca-Sp8
324	(GlcNAcb1-3)-3Galb-Sp8
325	(GlcNAcb1-4)(GlcNAcb1-6)-3,4,6GlcNAc
326	(GlcNAcb1-4)(GlcNAcb1-6)-4,6GalNAca
327	(GlcNAcb1-4)-(4Galb1-4)-4GlcNAcb
328	(GlcNAcb1-4)-(4GlcNAcb1-4)-4GlcNAcb
329	(GlcNAcb1-4)-3,4,6GlcNAc-Sp8
330	(GlcNAcb1-4)-4,6GalNAca-Sp8
331	(GlcNAcb1-4)-4GlcNAcb1-Sp8
332	(GlcNAcb1-6)-(3,6Galb1-4)-4GlcNAcb
333	(GlcNAcb1-6)-(6Galb1-4)-4GlcNAcb
334	(GlcNAcb1-6)-3,4,6GlcNAc-Sp8
335	(GlcNAcb1-6)-3,6GalNAca-Sp8
336	(GlcNAcb1-6)-3,6GlcNAca-Sp8
337	(GlcNAcb1-6)-4,6GalNAca-Sp8
338	(GlcNAcb1-6)-6GalNAca-Sp8
339	(Glca1-4)-4Glca-Sp8
340	(Glca1-4)-4Glcb-Sp8

341	(Glca1-6)-(6Glca1-6)-6Glcb
342	(Glcb1-4)-4Glcb-Sp8
343	(Glcb1-6)-6Glcb-Sp8
344	(KDNa2-3)-(3Galb1-3)-3GlcNAcb
345	(KDNa2-3)-(3Galb1-4)-4GlcNAcb
346	(Mana1-2)-(2Mana1-2)-2Mana1
347	(Mana1-2)-(2Mana1-3)-3,6Mana
348	(Mana1-2)-(2Mana1-3)-3,6Mana1
349	(Mana1-2)-(2Mana1-3)-3,6Manb1
350	(Mana1-2)-(2Mana1-3)-3Mana
351	(Mana1-2)-(2Mana1-6)-3,6Mana
352	(Mana1-2)-(2Mana1-6)-3,6Mana1
353	(Mana1-3)(2Mana1-6)-3,6Mana
354	(Mana1-3)(2Mana1-6)-3,6Mana1
355	(Mana1-3)(3,6Mana1-6)-3,6Manb1
356	(Mana1-3)(Mana1-6)-3,6Mana
357	(Mana1-3)(Mana1-6)-3,6Mana1
358	(Mana1-3)(Mana1-6)-3,6Manb1
359	(Mana1-3)-(3,6Mana1-6)-3,6Manb1
360	(Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
361	(Mana1-3)-3,6Mana-Sp9
362	(Mana1-6)-(3,6Mana1-6)-3,6Manb1
363	(Mana1-6)-(3,6Manb1-4)-4GlcNAcb1
364	(Mana1-6)-3,6Mana-Sp9
365	(Manb1-4)-4GlcNAcb-Sp0
366	(Neu5Aca2-3)(3GalNAcb1-4)-3,4Galb1
367	(Neu5Aca2-3)(GalNAcb1-4)-3,4Galb1
368	(Neu5Aca2-3)(Neu5Aca2-6)-3,6GalNAca
369	(Neu5Aca2-3)-(3(6-O-Su)Galb1-4)-3,4GlcNAcb
370	(Neu5Aca2-3)-(3(6OSO3)Galb1-4)-4GlcNAcb
371	(Neu5Aca2-3)-(3,4Galb1-4)-4GlcNAcb
372	(Neu5Aca2-3)-(3,4Galb1-4)-4Glcb
373	(Neu5Aca2-3)-(3GalNAcb1-4)-4GlcNAcb
374	(Neu5Aca2-3)-(3Galb1-3)-3(6OSO3)GalNAca
375	(Neu5Aca2-3)-(3Galb1-3)-3(6OSO3)GlcNAcb
376	(Neu5Aca2-3)-(3Galb1-3)-3,4GlcNAcb
377	(Neu5Aca2-3)-(3Galb1-3)-3,4GlcNAcb1
378	(Neu5Aca2-3)-(3Galb1-3)-3,6GalNAcb
379	(Neu5Aca2-3)-(3Galb1-3)-3GalNAca
380	(Neu5Aca2-3)-(3Galb1-3)-3GalNAcb1
381	(Neu5Aca2-3)-(3Galb1-3)-3GlcNAcb
382	(Neu5Aca2-3)-(3Galb1-3)-3GlcNAcb1
383	(Neu5Aca2-3)-(3Galb1-4)-3,4(6OSO3)GlcNAcb
384	(Neu5Aca2-3)-(3Galb1-4)-3,4GlcNAcb
385	(Neu5Aca2-3)-(3Galb1-4)-3,4GlcNAcb1
386	(Neu5Aca2-3)-(3Galb1-4)-4(6OSO3)GlcNAcb
387	(Neu5Aca2-3)-(3Galb1-4)-4GlcNAcb
388	(Neu5Aca2-3)-(3Galb1-4)-4GlcNAcb1
389	(Neu5Aca2-3)-(3Galb1-4)-4Glcb

390	(Neu5Aca2-3)-3,6GalNAca-Sp8
391	(Neu5Aca2-3)-3GalNAca-Sp8
392	(Neu5Aca2-3)-3Galb-Sp8
393	(Neu5Aca2-6)-(3,6GlcNAcb1-4)-4Galb1
394	(Neu5Aca2-6)-(6GalNAcb1-4)-4GlcNAcb
395	(Neu5Aca2-6)-(6Galb1-4)-4(6OSO3)GlcNAcb
396	(Neu5Aca2-6)-(6Galb1-4)-4GlcNAcb
397	(Neu5Aca2-6)-(6Galb1-4)-4GlcNAcb1
398	(Neu5Aca2-6)-(6Galb1-4)-4Glcb
399	(Neu5Aca2-6)-3,6GalNAca-Sp8
400	(Neu5Aca2-6)-3,6GalNAcb-Sp8
401	(Neu5Aca2-6)-3,6GlcNAca-Sp8
402	(Neu5Aca2-6)-6GalNAca-Sp8
403	(Neu5Aca2-6)-6Galb-Sp8
404	(Neu5Aca2-8)-(8Neu5Aca2-3)-3,4Galb1
405	(Neu5Aca2-8)-(8Neu5Aca2-3)-3Galb1
406	(Neu5Aca2-8)-8Neu5Aca-Sp8
407	(Neu5Acb2-6)-(6Galb1-4)-4GlcNAcb
408	(Neu5Acb2-6)-3,6GalNAca-Sp8
409	(Neu5Acb2-6)-3,6GlcNAca-Sp8
410	(Neu5Acb2-6)-6GalNAca-Sp8
411	(Neu5Gca2-3)-(3Galb1-3)-3,4GlcNAcb
412	(Neu5Gca2-3)-(3Galb1-3)-3GlcNAcb
413	(Neu5Gca2-3)-(3Galb1-4)-3,4GlcNAcb
414	(Neu5Gca2-3)-(3Galb1-4)-4GlcNAcb
415	(Neu5Gca2-3)-(3Galb1-4)-4Glcb
416	(Neu5Gca2-6)-(6Galb1-4)-4GlcNAcb
417	(Neu5Gca2-6)-6GalNAca-Sp0

Table A-IV. CFG array v2.0 tetra-saccharide Features Summary

Number of Features	tetra-saccharide
0	((2Galb1-3)-3GalNAcb1-4)(Neu5Aca2-3)-3,4Galb1
1	((2Mana1-2)-2Mana1-3)(3,6Mana1-6)-3,6Manb1
2	((2Mana1-2)-2Mana1-6)(Mana1-3)-3,6Mana
3	((2Mana1-3)-3,6Mana1-6)(2Mana1-3)-3,6Manb1
4	((2Mana1-6)-3,6Mana1-6)(2Mana1-3)-3,6Manb1
5	((3,4GalNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb
6	((3,4GalNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb1
7	((3,4GlcNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GalNAcb
8	((3,4GlcNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb
9	((3,4GlcNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb1
10	((3,4GlcNAcb1-4)-4Galb1-4)(Fuca1-3)-3,4GlcNAcb
11	((3,4GlcNAcb1-4)-4Galb1-4)(Fuca1-3)-3,4GlcNAcb1
12	((3Galb1-3)-3GalNAcb1-4)(Neu5Aca2-3)-3,4Galb1
13	((3OSO3)Galb1-3)(Fuca1-4)-3,4GlcNAcb-Sp8
14	((4GlcNAcb1-2)-2Mana1-3)(2Mana1-6)-3,6Manb1
15	((4GlcNAcb1-2)-2Mana1-6)(2Mana1-3)-3,6Manb1
16	((4GlcNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb
17	((4GlcNAcb1-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb1
18	((8Neu5Aca2-8)-8Neu5Aca2-3)(GalNAcb1-4)-3,4Galb1
19	((Fuca1-2)-2,3Galb1-4)(Fuca1-3)-3,4GlcNAcb
20	((Fuca1-2)-2Galb1-3)(Fuca1-4)-3,4GlcNAcb
21	((Fuca1-2)-2Galb1-4)(Fuca1-3)-3,4GalNAcb1
22	((Fuca1-2)-2Galb1-4)(Fuca1-3)-3,4GlcNAcb
23	((GalNAca1-3)-2,3Galb1-4)(Fuca1-3)-3,4GlcNAcb
24	((Gala1-3)-2,3Galb1-4)(Fuca1-3)-3,4GlcNAcb
25	((Gala1-3)-3Galb1-4)(Fuca1-3)-3,4GalNAcb
26	((Galb1-3)-3GalNAcb1-4)(Neu5Aca2-3)-3,4Galb1
27	((Galb1-4)-4GalNAca1-3)(Fuca1-2)-2,3Galb1
28	((Galb1-4)-4GalNAcb1-3)(Fuca1-2)-2,3Galb1
29	((Galb1-4)-4GlcNAcb1-3)(4GlcNAcb1-6)-3,6GalNAca
30	((Galb1-4)-4GlcNAcb1-6)(4GlcNAcb1-3)-3,6GalNAca
31	((Galb1-4)-4GlcNAcb1-6)(Galb1-3)-3,6GalNAca
32	((Galb1-4)-4GlcNAcb1-6)(Galb1-3)-3,6GlcNAca
33	((GlcNAcb1-2)-2Mana1-3)(2Mana1-6)-3,6Manb1
34	((GlcNAcb1-2)-2Mana1-6)(2Mana1-3)-3,6Manb1
35	((Mana1-2)-2Mana1-3)(2Mana1-6)-3,6Mana
36	((Mana1-2)-2Mana1-3)(2Mana1-6)-3,6Mana1
37	((Mana1-2)-2Mana1-3)(3,6Mana1-6)-3,6Manb1
38	((Mana1-2)-2Mana1-3)(Mana1-6)-3,6Mana1
39	((Mana1-2)-2Mana1-6)(2Mana1-3)-3,6Mana
40	((Mana1-2)-2Mana1-6)(2Mana1-3)-3,6Mana1
41	((Mana1-2)-2Mana1-6)(Mana1-3)-3,6Mana1
42	((Mana1-3)-3,6Mana1-6)(2Mana1-3)-3,6Manb1
43	((Mana1-3)-3,6Mana1-6)(Mana1-3)-3,6Manb1
44	((Mana1-6)-3,6Mana1-6)(2Mana1-3)-3,6Manb1
45	((Mana1-6)-3,6Mana1-6)(Mana1-3)-3,6Manb1
46	((Neu5Aca2-3)-3(6-O-Su)Galb1-4)(Fuca1-3)-3,4GlcNAcb

47	((Neu5Aca2-3)-3Galb1-3)(3Galb1-4)-3,4GlcNAcb
48	((Neu5Aca2-3)-3Galb1-3)(Fuca1-4)-3,4GlcNAcb
49	((Neu5Aca2-3)-3Galb1-3)(Fuca1-4)-3,4GlcNAcb1
50	((Neu5Aca2-3)-3Galb1-3)(Neu5Aca2-6)-3,6GalNAcb
51	((Neu5Aca2-3)-3Galb1-4)(3Galb1-3)-3,4GlcNAcb
52	((Neu5Aca2-3)-3Galb1-4)(Fuca1-3)-3,4(6OSO3)GlcNAcb
53	((Neu5Aca2-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb
54	((Neu5Aca2-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb1
55	((Neu5Aca2-8)-8Neu5Aca2-3)(GalNAcb1-4)-3,4Galb1
56	((Neu5Gca2-3)-3Galb1-3)(Fuca1-4)-3,4GlcNAcb
57	((Neu5Gca2-3)-3Galb1-4)(Fuca1-3)-3,4GlcNAcb
58	(2Galb1-3)(Fuca1-4)-3,4GlcNAcb-Sp8
59	(2Galb1-3)-(3GalNAcb1-3)-(3Gala1-4)4Galb1
60	(2Galb1-3)-(3GalNAcb1-3)-(3Galb1-4)4Glcb
61	(2Galb1-3)-(3GalNAcb1-3)-3Gala-Sp9
62	(2Galb1-4)(Fuca1-3)-(3,4GalNAcb1-3)-3Galb1
63	(2Galb1-4)-(3,4GalNAcb1-3)-(3Galb1-4)3,4GlcNAcb
64	(2Galb1-4)-(3,4GalNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
65	(2Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
66	(2Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb1
67	(2Mana1-2)-(2Mana1-3)-3Mana-Sp9
68	(2Mana1-2)-(2Mana1-6)-3,6Mana-Sp9
69	(2Mana1-3)(2Mana1-6)-3,6Mana-Sp9
70	(2Mana1-6)(2Mana1-3)-(3,6Mana1-6)-3,6Manb1
71	(2Mana1-6)(2Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
72	(2Mana1-6)(Mana1-3)-(3,6Mana1-6)-3,6Manb1
73	(2Mana1-6)-(3,6Mana1-6)-(3,6Manb1-4)4GlcNAcb1
74	(2Mana1-6)-(3,6Manb1-4)-(4GlcNAcb1-4)4GlcNAcb
75	(3,4GalNAcb1-3)-(3Galb1-4)-3,4GlcNAcb-Sp0
76	(3,4GlcNAcb1-3)-(3Galb1-4)-3,4GalNAcb-Sp0
77	(3,4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb-Sp0
78	(3,4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp0
79	(3,4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp8
80	(3,4GlcNAcb1-4)-(4Galb1-4)-(3,4GlcNAcb1-4)4Galb1
81	(3,4GlcNAcb1-4)-(4Galb1-4)-3,4GlcNAcb-Sp0
82	(3,6GlcNAcb1-4)-(4Galb1-4)-4Glcb-Sp10
83	(3,6Mana1-6)(2Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
84	(3,6Mana1-6)(Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
85	(3,6Mana1-6)-(3,6Manb1-4)-(4GlcNAcb1-4)4GlcNAcb
86	(3,6Manb1-4)-(4GlcNAcb1-4)-4GlcNAcb-Asn
87	(3,6Manb1-4)-(4GlcNAcb1-4)-4GlcNAcb-Gly
88	(3,6Manb1-4)-(4GlcNAcb1-4)-4GlcNAcb-Sp8
89	(3GalNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp0
90	(3GalNAcb1-3)-(3Galb1-4)-4Glcb-Sp10
91	(3GalNAcb1-3)-(3Galb1-4)-4Glcb-Sp8
92	(3GalNAcb1-4)(Neu5Aca2-3)-(3,4Galb1-4)-4Glcb
93	(3GalNAcb1-4)-(3,4Galb1-4)-4Glcb-Sp0
94	(3GalNAcb1-4)-(3,4Galb1-4)-4Glcb-Sp9
95	(3GalNAcb1-4)-(4Galb1-4)-4Glcb-Sp8

96	(3Gala1-4)-(4Galb1-4)-4GlcNAcb-Sp0
97	(3Gala1-4)-(4Galb1-4)-4Glca-Sp9
98	(3Gala1-4)-(4Galb1-4)-4Glcb-Sp0
99	(3Galb1-3)(3Galb1-4)-3,4GlcNAcb-Sp8
100	(3Galb1-3)(Fuca1-4)-3,4GlcNAcb-Sp0
101	(3Galb1-3)(Fuca1-4)-3,4GlcNAcb-Sp8
102	(3Galb1-3)(Neu5Aca2-6)-3,6GalNAcb-Sp8
103	(3Galb1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
104	(3Galb1-3)-(3GalNAcb1-3)-(3Gala1-4)4Galb1
105	(3Galb1-3)-(3GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
106	(3Galb1-4)(Fuca1-3)-(3,4GlcNAcb1-3)-3Galb
107	(3Galb1-4)(Fuca1-3)-(3,4GlcNAcb1-3)-3Galb1
108	(3Galb1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GalNAcb
109	(3Galb1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
110	(3Galb1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
111	(3Galb1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
112	(3Galb1-4)-(3,4GlcNAcb1-3)-3Galb-Sp8
113	(3Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
114	(3Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
115	(3Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb1
116	(3GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp0
117	(4GalNAca1-3)(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb
118	(4GalNAca1-3)-(2,3Galb1-4)-4GlcNAcb-Sp8
119	(4GalNAcb1-3)(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb
120	(4GalNAcb1-3)-(2,3Galb1-4)-4GlcNAcb-Sp8
121	(4Galb1-4)(Fuca1-3)-(3,4GlcNAcb1-4)-4Galb1
122	(4Galb1-4)-(3,4GlcNAcb1-4)-(4Galb1-4)3,4GlcNAcb
123	(4GlcNAcb1-3)(4GlcNAcb1-6)-3,6GalNAca-Sp8
124	(4GlcNAcb1-3)-(3Galb1-4)-3,4GlcNAcb-Sp0
125	(4GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp0
126	(4GlcNAcb1-3)-(3Galb1-4)-4Glcb-Sp0
127	(4GlcNAcb1-3)-(3Galb1-4)-4Glcb-Sp8
128	(6Galb1-4)-(4GlcNAcb1-2)-(2Mana1-3)3,6Manb1
129	(6Galb1-4)-(4GlcNAcb1-2)-(2Mana1-6)3,6Manb1
130	(6Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
131	(6Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
132	(8Neu5Aca2-3)-(3,4Galb1-4)-4Glcb-Sp0
133	(8Neu5Aca2-3)-(3Galb1-4)-4Glcb-Sp0
134	(8Neu5Aca2-8)-(8Neu5Aca2-3)-(3,4Galb1-4)4Glcb
135	(8Neu5Aca2-8)-(8Neu5Aca2-3)-(3Galb1-4)4Glcb
136	(9NAcNeu5Aca2-6)-(6Galb1-4)-4GlcNAcb-Sp8
137	(Fuca1-2)(GalNAca1-3)-2,3Galb-Sp8
138	(Fuca1-2)(GalNAcb1-3)-2,3Galb-Sp8
139	(Fuca1-2)(Gala1-3)-2,3Galb-Sp8
140	(Fuca1-2)-(2,3Galb1-3)-3GlcNAcb-Sp0
141	(Fuca1-2)-(2,3Galb1-4)-3,4GlcNAcb-Sp0
142	(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb-Sp0
143	(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb-Sp8
144	(Fuca1-2)-(2,3Galb1-4)-4Glcb-Sp0

145	(Fuca1-2)-(2,4Galb1-4)-4GalNAcb-Sp8
146	(Fuca1-2)-(2,4Galb1-4)-4GlcNAcb-Sp8
147	(Fuca1-2)-(2Galb1-3)-3,4GlcNAcb-Sp8
148	(Fuca1-2)-(2Galb1-3)-3GalNAca-Sp8
149	(Fuca1-2)-(2Galb1-3)-3GalNAcb-Sp0
150	(Fuca1-2)-(2Galb1-3)-3GalNAcb-Sp8
151	(Fuca1-2)-(2Galb1-4)-3,4GlcNAcb-Sp0
152	(Fuca1-2)-(2Galb1-4)-3,4GlcNAcb-Sp8
153	(Fuca1-2)-(2Galb1-4)-4GlcNAcb-Sp0
154	(Fuca1-2)-(2Galb1-4)-4GlcNAcb-Sp8
155	(Fuca1-2)-(2Galb1-4)-4Glcb-Sp0
156	(Fuca1-3)((3OSO3)Galb1-4)-3,4GlcNAcb-Sp8
157	(Fuca1-3)(2,3Galb1-4)-3,4GlcNAcb-Sp0
158	(Fuca1-3)(2Galb1-4)-3,4GlcNAcb-Sp0
159	(Fuca1-3)(2Galb1-4)-3,4GlcNAcb-Sp8
160	(Fuca1-3)(3(6-O-Su)Galb1-4)-3,4GlcNAcb-Sp8
161	(Fuca1-3)(3Galb1-4)-3,4(6OSO3)GlcNAcb-Sp8
162	(Fuca1-3)(3Galb1-4)-3,4GalNAcb-Sp0
163	(Fuca1-3)(3Galb1-4)-3,4GalNAcb-Sp8
164	(Fuca1-3)(3Galb1-4)-3,4GlcNAcb-Sp0
165	(Fuca1-3)(3Galb1-4)-3,4GlcNAcb-Sp8
166	(Fuca1-3)(4Galb1-4)-3,4GlcNAcb-Sp0
167	(Fuca1-3)(GalNAcb1-4)-3,4GlcNAcb-Sp0
168	(Fuca1-3)(Galb1-4)-3,4GlcNAcb-Sp0
169	(Fuca1-3)(Galb1-4)-3,4GlcNAcb-Sp8
170	(Fuca1-3)-(3,4GalNAcb1-3)-(3Galb1-4)3,4GlcNAcb
171	(Fuca1-3)-(3,4GalNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
172	(Fuca1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GalNAcb
173	(Fuca1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
174	(Fuca1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
175	(Fuca1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
176	(Fuca1-3)-(3,4GlcNAcb1-3)-3Galb-Sp8
177	(Fuca1-4)(3Galb1-3)-(3,4GlcNAcb1-3)-3Galb1
178	(Fuca1-4)(Galb1-3)-(3,4GlcNAcb1-3)-3Galb1
179	(Fuca1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
180	(Fuca1-4)-(3,4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
181	(GalNAca1-3)(Fuca1-2)-(2,3Galb1-3)-3GlcNAcb
182	(GalNAca1-3)(Fuca1-2)-(2,3Galb1-4)-3,4GlcNAcb
183	(GalNAca1-3)(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb
184	(GalNAca1-3)(Fuca1-2)-(2,3Galb1-4)-4Glcb
185	(GalNAca1-3)-(2,3Galb1-3)-3GlcNAcb-Sp0
186	(GalNAca1-3)-(2,3Galb1-4)-3,4GlcNAcb-Sp0
187	(GalNAca1-3)-(2,3Galb1-4)-4GlcNAcb-Sp0
188	(GalNAca1-3)-(2,3Galb1-4)-4GlcNAcb-Sp8
189	(GalNAca1-3)-(2,3Galb1-4)-4Glcb-Sp0
190	(GalNAca1-4)(Fuca1-2)-(2,4Galb1-4)-4GlcNAcb
191	(GalNAca1-4)-(2,4Galb1-4)-4GlcNAcb-Sp8
192	(GalNAcb1-4)(8Neu5Aca2-3)-(3,4Galb1-4)-4Glcb
193	(GalNAcb1-4)(Neu5Aca2-3)-(3,4Galb1-4)-4GlcNAcb

194	(GalNAcb1-4)(Neu5Aca2-3)-(3,4Galb1-4)-4Glcb
195	(GalNAcb1-4)-(3,4Galb1-4)-4GlcNAcb-Sp0
196	(GalNAcb1-4)-(3,4Galb1-4)-4GlcNAcb-Sp8
197	(GalNAcb1-4)-(3,4Galb1-4)-4Glcb-Sp0
198	(Gala1-3)(Fuca1-2)-(2,3Galb1-3)-3GlcNAcb
199	(Gala1-3)(Fuca1-2)-(2,3Galb1-4)-3,4GlcNAcb
200	(Gala1-3)(Fuca1-2)-(2,3Galb1-4)-4GlcNAcb
201	(Gala1-3)(Fuca1-2)-(2,3Galb1-4)-4Glcb
202	(Gala1-3)-(2,3Galb1-3)-3GlcNAcb-Sp0
203	(Gala1-3)-(2,3Galb1-4)-3,4GlcNAcb-Sp0
204	(Gala1-3)-(2,3Galb1-4)-4GlcNAcb-Sp0
205	(Gala1-3)-(2,3Galb1-4)-4Glcb-Sp0
206	(Gala1-3)-(3,4Galb1-4)-4GlcNAcb-Sp8
207	(Gala1-3)-(3Galb1-3)-3GlcNAcb-Sp0
208	(Gala1-3)-(3Galb1-4)-3,4GalNAcb-Sp8
209	(Gala1-3)-(3Galb1-4)-4GlcNAcb-Sp8
210	(Gala1-3)-(3Galb1-4)-4Glcb-Sp0
211	(Gala1-4)(Fuca1-2)-(2,4Galb1-4)-4GalNAcb
212	(Gala1-4)(Gala1-3)-(3,4Galb1-4)-4GlcNAcb
213	(Gala1-4)-(2,4Galb1-4)-4GalNAcb-Sp8
214	(Gala1-4)-(3,4Galb1-4)-4GlcNAcb-Sp8
215	(Gala1-4)-(4Galb1-4)-4GalNAcb-Sp0
216	(Gala1-4)-(4Galb1-4)-4GalNAcb-Sp8
217	(Gala1-4)-(4Galb1-4)-4Galb-Sp0
218	(Galb1-3)(4GlcNAcb1-6)-3,6GalNAca-Sp8
219	(Galb1-3)(4GlcNAcb1-6)-3,6GlcNAca-Sp8
220	(Galb1-3)(Fuca1-4)-3,4GlcNAc-Sp0
221	(Galb1-3)(Fuca1-4)-3,4GlcNAc-Sp8
222	(Galb1-3)(Fuca1-4)-3,4GlcNAcb-Sp8
223	(Galb1-3)(GlcNAcb1-6)-3,6GalNAca-Sp8
224	(Galb1-3)(GlcNAcb1-6)-3,6GlcNAca-Sp8
225	(Galb1-3)(Neu5Aca2-6)-3,6GalNAca-Sp8
226	(Galb1-3)(Neu5Aca2-6)-3,6GlcNAca-Sp8
227	(Galb1-3)(Neu5Acb2-6)-3,6GalNAca-Sp8
228	(Galb1-3)(Neu5Acb2-6)-3,6GlcNAca-Sp8
229	(Galb1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb
230	(Galb1-3)-(3,4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb
231	(Galb1-3)-(3GalNAcb1-3)-(3Gala1-4)4Galb1
232	(Galb1-3)-(3GalNAcb1-3)-(3Galb1-4)4GlcNAcb
233	(Galb1-3)-(3GalNAcb1-3)-(3Galb1-4)4Glcb
234	(Galb1-4)(Fuca1-3)-(3,4GlcNAcb1-4)-4Galb1
235	(Galb1-4)-(3,4GlcNAcb1-4)-(4Galb1-4)3,4GlcNAcb
236	(Galb1-4)-(3,4GlcNAcb1-4)-(4Galb1-4)3,4GlcNAcb1
237	(Galb1-4)-(4GalNAca1-3)-(2,3Galb1-4)4GlcNAcb
238	(Galb1-4)-(4GalNAcb1-3)-(2,3Galb1-4)4GlcNAcb
239	(Galb1-4)-(4GlcNAcb1-2)-(2Mana1-3)3,6Manb1
240	(Galb1-4)-(4GlcNAcb1-2)-(2Mana1-6)3,6Manb1
241	(Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)3,4GlcNAcb1
242	(Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb

243	(Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4GlcNAcb1
244	(Galb1-4)-(4GlcNAcb1-3)-(3Galb1-4)4Glcb
245	(Galb1-4)-(4GlcNAcb1-3)-3,6GalNAca-Sp8
246	(Galb1-4)-(4GlcNAcb1-3)-3GalNAca-Sp8
247	(Galb1-4)-(4GlcNAcb1-6)-3,6GalNAca-Sp8
248	(Galb1-4)-(4GlcNAcb1-6)-3,6GlcNAca-Sp8
249	(Galb1-4)-(4GlcNAcb1-6)-6GalNAca-Sp8
250	(GlcNAca1-3)-(3Galb1-4)-4GlcNAcb-Sp8
251	(GlcNAca1-6)-(6Galb1-4)-4GlcNAcb-Sp8
252	(GlcNAcb1-2)-(2Galb1-3)-3GlcNAca-Sp8
253	(GlcNAcb1-3)(GlcNAcb1-4)(GlcNAcb1-6)-3,4,6GlcNAc
254	(GlcNAcb1-3)(GlcNAcb1-4)-3,4,6GlcNAc-Sp8
255	(GlcNAcb1-3)(GlcNAcb1-6)-3,4,6GlcNAc-Sp8
256	(GlcNAcb1-3)(GlcNAcb1-6)-3,6GlcNAca-Sp8
257	(GlcNAcb1-3)-(3,6Galb1-4)-4GlcNAcb-Sp8
258	(GlcNAcb1-3)-(3Galb1-3)-3GalNAca-Sp8
259	(GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp0
260	(GlcNAcb1-3)-(3Galb1-4)-4GlcNAcb-Sp8
261	(GlcNAcb1-3)-(3Galb1-4)-4Glcb-Sp0
262	(GlcNAcb1-4)(GlcNAcb1-6)-3,4,6GlcNAc-Sp8
263	(GlcNAcb1-4)(GlcNAcb1-6)-4,6GalNAca-Sp8
264	(GlcNAcb1-4)-(4Galb1-4)-4GlcNAcb-Sp8
265	(GlcNAcb1-4)-(4GlcNAcb1-4)-4GlcNAcb-Sp8
266	(GlcNAcb1-6)(GlcNAcb1-3)-(3,6Galb1-4)-4GlcNAcb
267	(GlcNAcb1-6)-(3,6Galb1-4)-4GlcNAcb-Sp8
268	(GlcNAcb1-6)-(6Galb1-4)-4GlcNAcb-Sp8
269	(Glca1-6)-(6Glca1-6)-6Glcb-Sp8
270	(KDNa2-3)-(3Galb1-3)-3GlcNAcb-Sp0
271	(KDNa2-3)-(3Galb1-4)-4GlcNAcb-Sp0
272	(Mana1-2)-(2Mana1-2)-(2Mana1-3)3,6Manb1
273	(Mana1-2)-(2Mana1-2)-(2Mana1-3)3Mana
274	(Mana1-2)-(2Mana1-2)-(2Mana1-6)3,6Mana
275	(Mana1-2)-(2Mana1-3)-3,6Mana-Sp9
276	(Mana1-2)-(2Mana1-3)-3Mana-Sp9
277	(Mana1-2)-(2Mana1-6)-3,6Mana-Sp9
278	(Mana1-3)(2Mana1-6)-3,6Mana-Sp9
279	(Mana1-3)(Mana1-6)-3,6Mana-Sp9
280	(Mana1-6)(2Mana1-3)-(3,6Mana1-6)-3,6Manb1
281	(Mana1-6)(Mana1-3)-(3,6Mana1-6)-3,6Manb1
282	(Mana1-6)(Mana1-3)-(3,6Manb1-4)-4GlcNAcb1
283	(Mana1-6)-(3,6Mana1-6)-(3,6Manb1-4)4GlcNAcb1
284	(Mana1-6)-(3,6Manb1-4)-(4GlcNAcb1-4)4GlcNAcb
285	(Neu5Aca2-3)(Neu5Aca2-6)-3,6GalNAca-Sp8
286	(Neu5Aca2-3)-(3(6-O-Su)Galb1-4)-3,4GlcNAcb-Sp8
287	(Neu5Aca2-3)-(3(6OSO3)Galb1-4)-4GlcNAcb-Sp8
288	(Neu5Aca2-3)-(3,4Galb1-4)-4GlcNAcb-Sp0
289	(Neu5Aca2-3)-(3,4Galb1-4)-4GlcNAcb-Sp8
290	(Neu5Aca2-3)-(3,4Galb1-4)-4Glcb-Sp0
291	(Neu5Aca2-3)-(3,4Galb1-4)-4Glcb-Sp9

292	(Neu5Aca2-3)-(3GalNAcb1-4)-4GlcNAcb-Sp0
293	(Neu5Aca2-3)-(3Galb1-3)-(3,4GlcNAcb1-3)3Galb1
294	(Neu5Aca2-3)-(3Galb1-3)-(3GalNAcb1-3)3Gala1
295	(Neu5Aca2-3)-(3Galb1-3)-(3GalNAcb1-4)3,4Galb1
296	(Neu5Aca2-3)-(3Galb1-3)-(3GlcNAcb1-3)3Galb1
297	(Neu5Aca2-3)-(3Galb1-3)-3(6OSO3)GalNAca-Sp8
298	(Neu5Aca2-3)-(3Galb1-3)-3(6OSO3)GlcNAcb-Sp8
299	(Neu5Aca2-3)-(3Galb1-3)-3,4GlcNAcb-Sp8
300	(Neu5Aca2-3)-(3Galb1-3)-3,6GalNAcb-Sp8
301	(Neu5Aca2-3)-(3Galb1-3)-3GalNAca-Sp8
302	(Neu5Aca2-3)-(3Galb1-3)-3GlcNAcb-Sp0
303	(Neu5Aca2-3)-(3Galb1-3)-3GlcNAcb-Sp8
304	(Neu5Aca2-3)-(3Galb1-4)-3,4(6OSO3)GlcNAcb-Sp8
305	(Neu5Aca2-3)-(3Galb1-4)-3,4GlcNAcb-Sp0
306	(Neu5Aca2-3)-(3Galb1-4)-3,4GlcNAcb-Sp8
307	(Neu5Aca2-3)-(3Galb1-4)-4(6OSO3)GlcNAcb-Sp8
308	(Neu5Aca2-3)-(3Galb1-4)-4GlcNAcb-Sp0
309	(Neu5Aca2-3)-(3Galb1-4)-4GlcNAcb-Sp8
310	(Neu5Aca2-3)-(3Galb1-4)-4Glcb-Sp0
311	(Neu5Aca2-3)-(3Galb1-4)-4Glcb-Sp8
312	(Neu5Aca2-6)(Galb1-3)-(3,6GlcNAcb1-4)-4Galb1
313	(Neu5Aca2-6)-(3,6GlcNAcb1-4)-(4Galb1-4)4Glcb
314	(Neu5Aca2-6)-(6GalNAcb1-4)-4GlcNAcb-Sp0
315	(Neu5Aca2-6)-(6Galb1-4)-(4GlcNAcb1-2)2Mana1
316	(Neu5Aca2-6)-(6Galb1-4)-(4GlcNAcb1-3)3Galb1
317	(Neu5Aca2-6)-(6Galb1-4)-4(6OSO3)GlcNAcb-Sp8
318	(Neu5Aca2-6)-(6Galb1-4)-4GlcNAcb-Sp0
319	(Neu5Aca2-6)-(6Galb1-4)-4GlcNAcb-Sp8
320	(Neu5Aca2-6)-(6Galb1-4)-4Glcb-Sp0
321	(Neu5Aca2-6)-(6Galb1-4)-4Glcb-Sp8
322	(Neu5Aca2-8)-(8Neu5Aca2-3)-(3,4Galb1-4)4Glcb
323	(Neu5Aca2-8)-(8Neu5Aca2-3)-(3Galb1-4)4Glcb
324	(Neu5Aca2-8)-(8Neu5Aca2-8)-8Neu5Aca-Sp8
325	(Neu5Acb2-6)-(6Galb1-4)-4GlcNAcb-Sp8
326	(Neu5Gca2-3)-(3Galb1-3)-3,4GlcNAcb-Sp0
327	(Neu5Gca2-3)-(3Galb1-3)-3GlcNAcb-Sp0
328	(Neu5Gca2-3)-(3Galb1-4)-3,4GlcNAcb-Sp0
329	(Neu5Gca2-3)-(3Galb1-4)-4GlcNAcb-Sp0
330	(Neu5Gca2-3)-(3Galb1-4)-4Glcb-Sp0
331	(Neu5Gca2-6)-(6Galb1-4)-4GlcNAcb-Sp0

References

[1] B. T. S. Da Wei Huang and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature protocols,* vol. 4, pp. 44-57, 2008.

[2] D. Gilbert, "Pise: Software for building bioinformatics webs," *Briefings in bioinformatics,* vol. 3, p. 405, 2002.

[3] A. G. Rust, *et al.,* "Genome annotation techniques: new approaches and challenges," *Drug discovery today,* vol. 7, pp. S70-S76, 2002.

[4] D. Gilbert, "Bioinformatics software resources," *Briefings in bioinformatics,* vol. 5, p. 300, 2004.

[5] D. S. Roos, "Bioinformatics--trying to swim in a sea of data," *Science,* vol. 291, p. 1260, 2001.

[6] N. Siva, "1000 Genomes project," *Nature biotechnology,* vol. 26, pp. 256-256, 2008.

[7] D. J. Munroe and T. J. R. Harris, "Third-generation sequencing fireworks at Marco Island," *Nature biotechnology,* vol. 28, pp. 426-428, 2010.

[8] B. A. Flusberg, *et al.,* "Direct detection of DNA methylation during single-molecule, real-time sequencing," *Nature methods,* vol. 7, p. 461, 2010.

[9] P. H. Carns, *et al.,* "PVFS: A parallel file system for linux clusters," 2000, pp. 28-28.

[10] A. S. Foundation., "Hadoop Distributed File System," *http://hadoop.apache.org/hdfs/,* 2011.

[11] F. Schmuck and R. Haskin, "GPFS: A shared-disk file system for large computing clusters," 2002, pp. 231–244.

[12] P. Schwan, "Lustre: Building a file system for 1000-node clusters," 2003.

[13] A. Bayat, "Bioinformatics," *Bmj,* vol. 324, p. 1018, 2002.

[14] P. Barham, *et al.,* "Xen and the art of virtualization," 2003, pp. 164-177.

[15] I. Habib, "Virtualization with kvm," *Linux Journal,* vol. 2008, p. 8, 2008.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[17] R. E. Fan, *et al.*, "Working set selection using second order information for training support vector machines," *The Journal of Machine Learning Research,* vol. 6, pp. 1889-1918, 2005.

[18] T. Joachims, "Making large scale SVM learning practical," 1999.

[19] T. Kudoh, "Tinysvm: Support vector machines," *http://cl.aist-nara. ac. jp/~ taku-ku/software/TinySVM/index. html,* 2000.

[20] S. Wold, *et al.*, "PLS-regression: a basic tool of chemometrics," *Chemometrics and intelligent laboratory systems,* vol. 58, pp. 109-130, 2001.

[21] GMOD, "Overview," *http://gmod.org/wiki/Overview,* 2011.

[22] S. F. Altschul, *et al.*, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.

[23] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences,* vol. 85, p. 2444, 1988.

[24] J. Stajich and E. Birney, "The Bioperl project: motivation and usage," *ACM SIGBIO Newsletter,* vol. 20, pp. 13-14, 2000.

[25] J. K. VanDyk, *Pro Drupal Development*: Springer, 2008.

[26] S. D. Mooney and P. H. Baenziger, "Extensible open source content management systems and frameworks: a solution for many needs of a bioinformatics group," *Briefings in bioinformatics,* vol. 9, p. 69, 2008.

[27] S. P. Ficklin, *et al.*, "Tripal: a construction toolkit for online genome databases," *Database,* vol. 2011, 2011.

[28] C. J. Mungall and D. B. Emmert, "A Chado case study: an ontology-based modular schema for representing genome-associated biological information," *Bioinformatics,* vol. 23, p. i337, 2007.

[29] "Generic Model Organism Database (GMOD) Project ".

[30] S. Temnykh, *et al.*, "Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential," *Genome research,* vol. 11, p. 1441, 2001.

[31] B. N, "FLIP: a Unix C program used to find/translate orfs," *bionet software,* 1997.

[32] "Organelle Genome Megasequencing Project (OGMP), Biochemistry Department, University of Montreal," *http://www.bch.umontreal.ca/ogmp/manlinks/flip.txt*.

[33] S. Rozen and H. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers," *Methods Mol Biol,* vol. 132, pp. 365-386, 2000.

[34] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome research,* vol. 9, p. 868, 1999.

[35] J. D. Tisdall, *Beginning Perl for bioinformatics*: O'Reilly Media, 2001.

[36] S. Guelich, *et al.*, *CGI programming with Perl*: O'Reilly & Associates, Inc., 2000.

[37] G. Booch, "Object-oriented design," *ACM SIGAda Ada Letters,* vol. 1, pp. 64-76, 1982.

[38] M. Rubel, "Rsnapshot: A Remote Filesystem Snapshot Utility Based on Rsync. 2005," *URL http://rsnapshot. org*.

[39] J. L. Ledford, *et al.*, *Google analytics*: Wiley, 2010.

[40] "rsync," *http://rsync.samba.org/,* 2011.

[41] Z. J. Chen, *et al.*, "Toward sequencing cotton (Gossypium) genomes," *Plant physiology,* vol. 145, p. 1303, 2007.

[42] J. A. Udall, *et al.*, "A global assembly of cotton ESTs," *Genome research,* vol. 16, p. 441, 2006.

[43] H. B. Zhang, *et al.*, "Recent advances in cotton genomics," *Int J Plant Genomics,* vol. 742304, 2008.

[44] A. Blenda, *et al.*, "CMD: a cotton microsatellite database resource for Gossypium genomics," *BMC genomics,* vol. 7, p. 132, 2006.

[45] E. Taliercio, *et al.*, "Analysis of ESTs from multiple Gossypium hirsutum tissues and identification of SSRs," *Genome,* vol. 49, pp. 306-319, 2006.

[46] J. Xiao, *et al.*, "New SSR Markers for Use in Cotton (Gossypium spp.) Improvement," 2009.

[47] S. Hoffman, *et al.*, "Identification of 700 new microsatellite loci from cotton (G. hirsutum L.)," 2007.

[48] L. L. Tu, *et al.*, "Genes expression analyses of sea-island cotton (Gossypium barbadense L.) during fiber development," *Plant cell reports,* vol. 26, pp. 1309-1320, 2007.

[49] T. S. T. Schwarzacher, *et al.*, "Characteristics and analysis of simple sequence repeats in the cotton genome based on a linkage map constructed from a BC1 population between Gossypium hirsutum and G. barbadense," *Genome,* vol. 51, pp. 534-546, 2008.

[50] J. Rong, *et al.*, "A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium)," *Genetics,* vol. 166, p. 389, 2004.

[51] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001.

[52] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics,* vol. 22, pp. 1658-1659, 2006.

[53] K. Youens-Clark, *et al.*, "CMap 1.01: a comparative mapping application for the Internet," *Bioinformatics,* vol. 25, p. 3040, 2009.

[54] A. Varki, *Essentials of glycobiology*: Cold Spring Harbor Laboratory Pr, 1999.

[55] D. H. Dube and C. R. Bertozzi, "Glycans in cancer and inflammation - potential for therapeutics and diagnostics," *Nat Rev Drug Discov,* vol. 4, pp. 477-488, 2005.

[56] M. M. Fuster and J. D. Esko, "The sweet and sour of cancer: glycans as novel therapeutic targets," *Nat Rev Cancer,* vol. 5, pp. 526-542, 2005.

[57] K. S. Lau and J. W. Dennis, "N-Glycans in cancer progression," *Glycobiology,* vol. 18, pp. 750-760, October 1, 2008 2008.

[58] N. Perrimon and M. Bernfield, "Specificities of heparan sulphate proteoglycans in developmental processes," *Nature,* vol. 404, pp. 725-728, 2000.

[59] X. Lin, "Functions of heparan sulfate proteoglycans in cell signaling during development," *Development,* vol. 131, pp. 6009-6021, December 15, 2004 2004.

[60] J. W. Dennis, *et al.*, "Protein glycosylation in development and disease," *BioEssays,* vol. 21, pp. 412-421, 1999.

[61] J. Stevens, *et al.*, "Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities," *Journal of molecular biology,* vol. 355, pp. 1143-1155, 2006.

[62] A. Alkhalil, *et al.*, "Structural Requirements for the Adherence ofPlasmodium falciparum-infected Erythrocytes to Chondroitin Sulfate Proteoglycans of Human Placenta," *Journal of Biological Chemistry,* vol. 275, pp. 40357-40364, December 22, 2000 2000.

[63] J. Liu, *et al.*, "Characterization of a Heparan Sulfate Octasaccharide That Binds to Herpes Simplex Virus Type 1 Glycoprotein D," *Journal of Biological Chemistry,* vol. 277, pp. 33456-33467, September 6, 2002 2002.

[64] A. Chandrasekaran, *et al.*, "Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin," *Nature biotechnology,* vol. 26, pp. 107-113, 2008.

[65] C. Y. Wu, *et al.*, "New development of glycan arrays," *Org. Biomol. Chem.,* vol. 7, pp. 2247-2254, 2009.

[66] O. Blixt, *et al.*, "Printed covalent glycan array for ligand profiling of diverse glycan binding proteins," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, p. 17033, 2004.

[67] M. E. Taylor and K. Drickamer, "Structural insights into what glycan arrays tell us about how glycan-binding proteins interact with their ligands," *Glycobiology,* vol. 19, p. 1155, 2009.

[68] K. Drickamer and M. E. Taylor, "Glycan arrays for functional glycomics," *Genome Biol,* vol. 3, p. 1034, 2002.

[69] A. Porter, *et al.*, "A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins," *Glycobiology,* vol. 20, p. 369, 2010.

[70] Y. Yamanishi, *et al.*, "Glycan classification with tree kernels," *Bioinformatics,* vol. 23, p. 1211, 2007.

[71] T. Kuboyama, *et al.*, "A gram distribution kernel applied to glycan classification and motif extraction," *Genome Informatics Series,* vol. 17, p. 25, 2006.

[72] I. G. Chong and C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and intelligent laboratory systems,* vol. 78, pp. 103-112, 2005.

[73] H. Yamamoto, *et al.*, "Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for enzyme gene involved in glycan biosynthesis," *PloS one,* vol. 2, p. e1232, 2007.

[74] Y. Naito, *et al.*, "Germinal center marker GL7 probes activation-dependent repression of N-glycolylneuraminic acid, a sialic acid species involved in the negative modulation of B-cell activation," *Molecular and cellular biology,* vol. 27, pp. 3008-22, Apr 2007.

[75] E. v. Liempt, *et al.*, "Specificity of DC-SIGN for mannose- and fucose-containing glycans," *FEBS letters,* vol. 580, pp. 6123-31, Nov 13 2006.