



Glycome Informatics: algorithms and tools

Kiyoko F. Aoki-Kinoshita

Bioinformatics Department

Faculty of Engineering

Soka University



Data mining for glycobiology

- Probabilistic models
- Kernel methods
- Text mining

- Applications of data mining

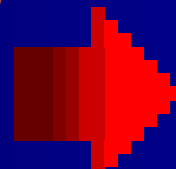
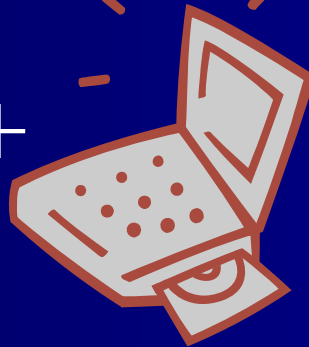


Machine Learning

Data samples



+



Rules, hypotheses,
models

New data samples



Prediction

Definition: the automatic extraction of rules and patterns from within large amounts of data

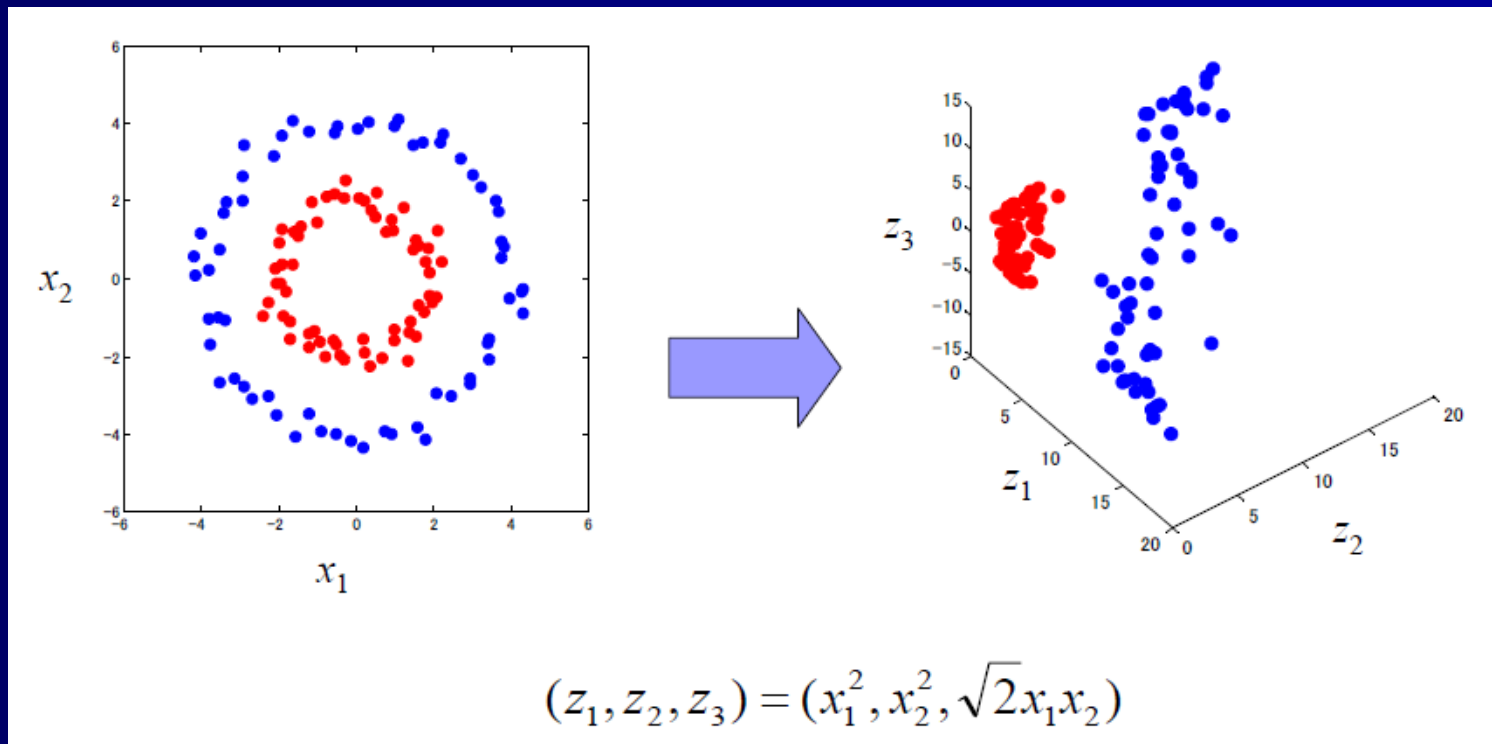


Machine Learning Methods

- PSTMM (Probabilistic Sibling-dependent Tree Markov Model)
 - Learns patterns from glycan structures
- Profile PSTMM
 - Extracts patterns (as profiles) from glycan structures
- Kernel methods
 - Classification of glycans
 - Extraction of “features” to predict glycan biomarkers

Kernel Method

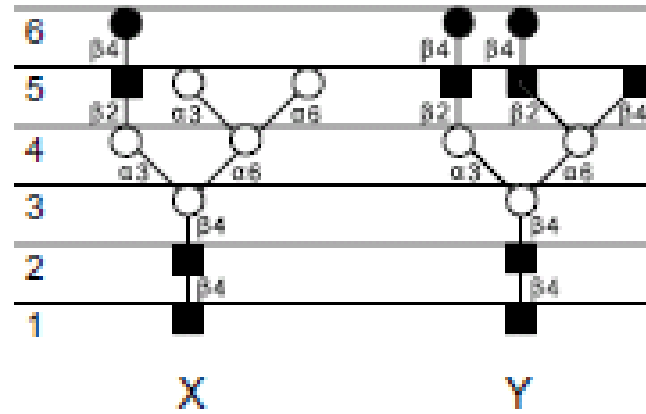
■ Classification method


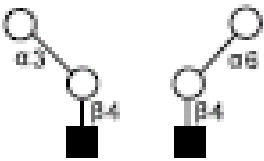
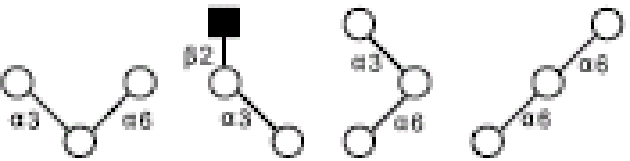
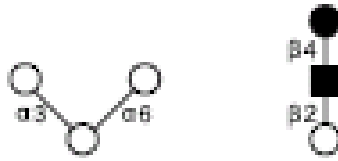

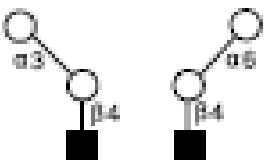
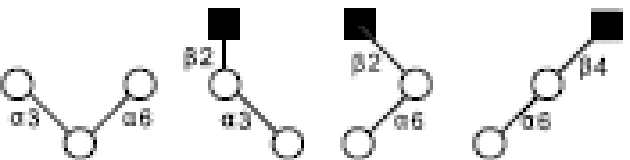
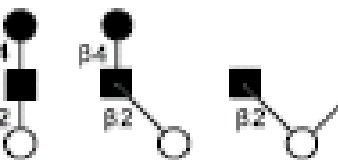




Leukemia kernel

- Extracted glycan structures from CarbBank
- Pre-analysis showed that the tri-saccharide structure was most effective for classification
- Furthermore, since the non-reducing end is usually the portion being recognized, this information was included in the kernel model



layer	1	2	3	4
X				
Y				



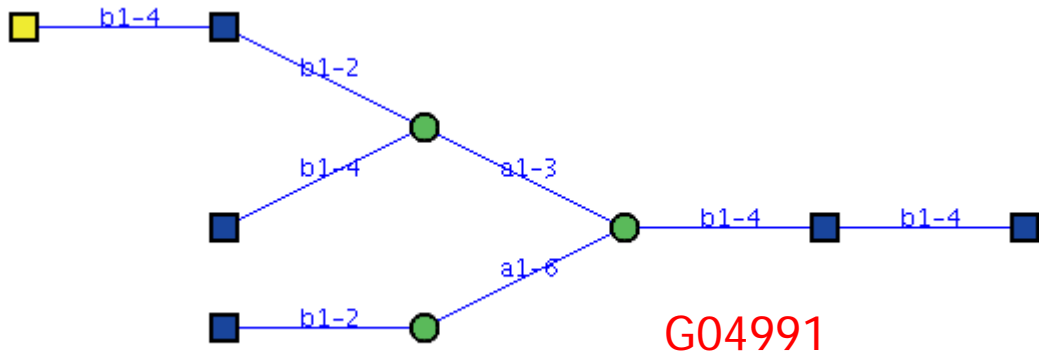
Predicted glycan markers for leukemic cells

Substructures	Layer	Scores
<i>Leukemic cells</i>		
α -D-Neup5Ac-(2→3)- β -D-Galp-(1→4)-D-GlcpNAc	5	161.2
β -D-Galp-(1→4)- β -D-GlcpNAc-(1→2)-D-Manp	4	159.6
α -D-Neup5Ac-(2→6)- β -D-Galp-(1→4)-D-GlcpNAc	5	148.8
β -D-GlcpNAc-(1→2)- α -D-Manp-(1→3)-D-Manp	3	78.7
β -D-GlcpNAc-(1→2)- α -D-Manp-(1→6)-D-Manp	3	77.6

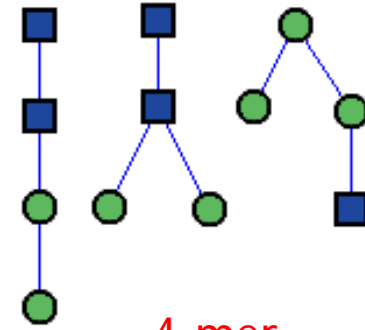


Other kernels

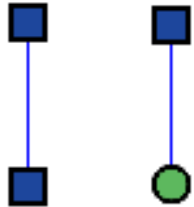
- Q-gram distribution kernel:
 - Wanted to be able to analyze any data regardless of marker structure or size
 - Definition of q-gram: A sub-tree containing q nodes
 - All of the q-grams for a particular glycan were included in the kernel
- Multiple kernel:
 - A kernel of kernels



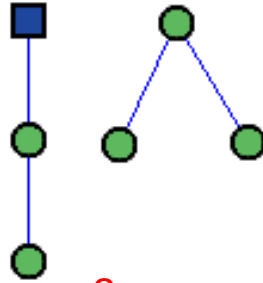
G04991



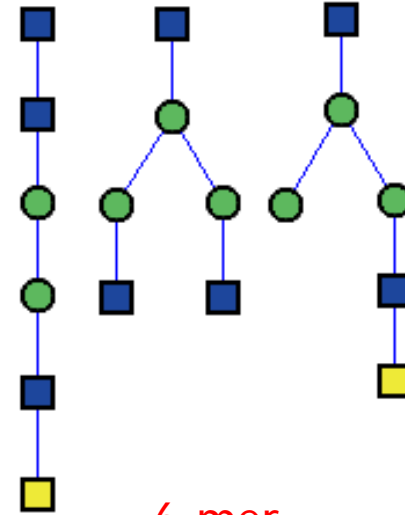
4-mer



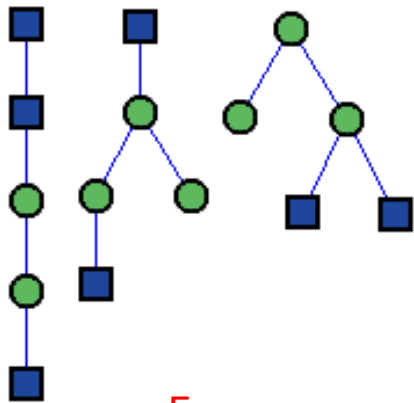
2-mer



3-mer



6-mer



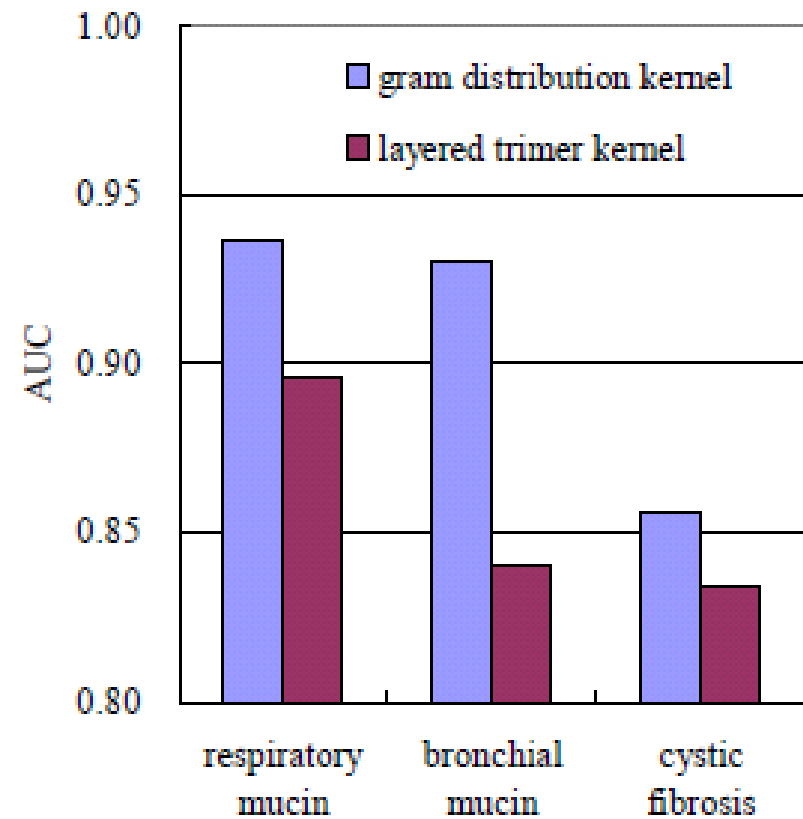
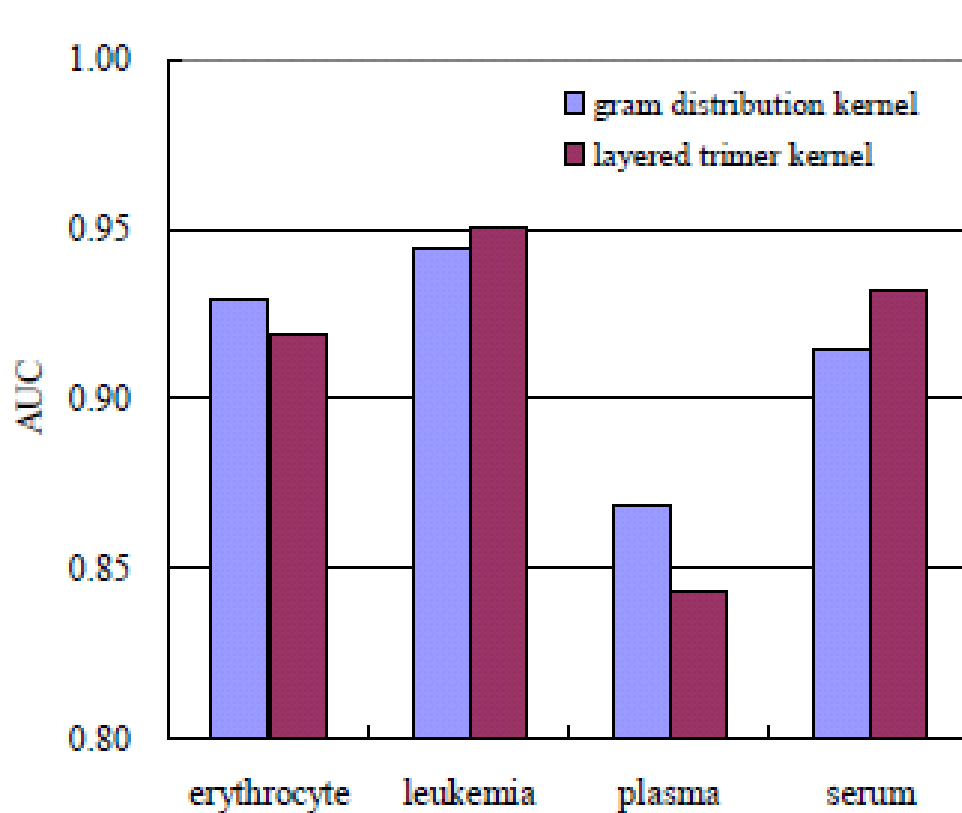
5-mer

q-grams



Results of gram distribution kernel

- Gram dist. vs. leukemia kernel





Summary of kernels

- Using a gram distribution, potential biomarkers of the appropriate size can be extracted from the data
- We are currently analyzing the various kernels to determine the most appropriate features for glycan data
- A web-based tool of the q-gram kernel that performs both training and classification is in the works



Applications of data mining to glycobiology

- Kernels can be utilized in many ways
 - Feature extract methods for detecting putative biomarkers
 - Cell-specific glycan structures can be extracted
 - Sequences of glycan binding proteins can be included in a new kernel to predict binding domains
 - Many more possibilities, depending on the data at hand

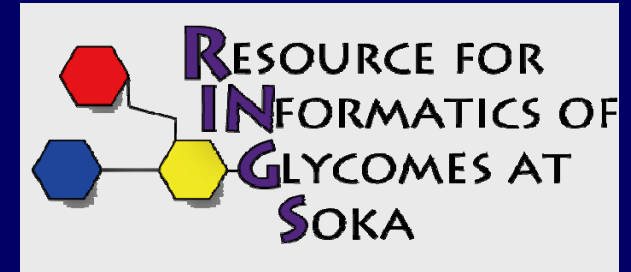


Resource for Glycome Informatics at Soka (RINGS)

- <http://rings.t.soka.ac.jp/>
- Motivation: a free resource for glycome informatics
- Started April, 2006
- Currently includes data from CAZy, KEGG and Glycosciences.de
 - Glycan structures, chemical reactions, glyco-related genes and proteins



RINGS: Tools



- BLAST search for glycan-related genes
- 2D glycan drawing and searching tool (DrawRINGS)
- Glycan score matrix generator
- ProfilePSTMM tool
- Gram distribution training and classification tools (Currently under construction)



Web services

- Defined as “a software system designed to support interoperable Machine-to-Machine interaction over a network”
- Computational tasks can be provided as “web services” which take specific data as input and return results in a particular format, without the need of a web browser
- That is, these are small tools that can be executed over the web from local software



RINGS Web services

- Given a particular glycan structure, retrieve all similar glycan structures in the database
 - Including similarity score
- Given a glycan in KCF format, return its corresponding LINUCS ID
 - Vice-versa
- Others coming soon...



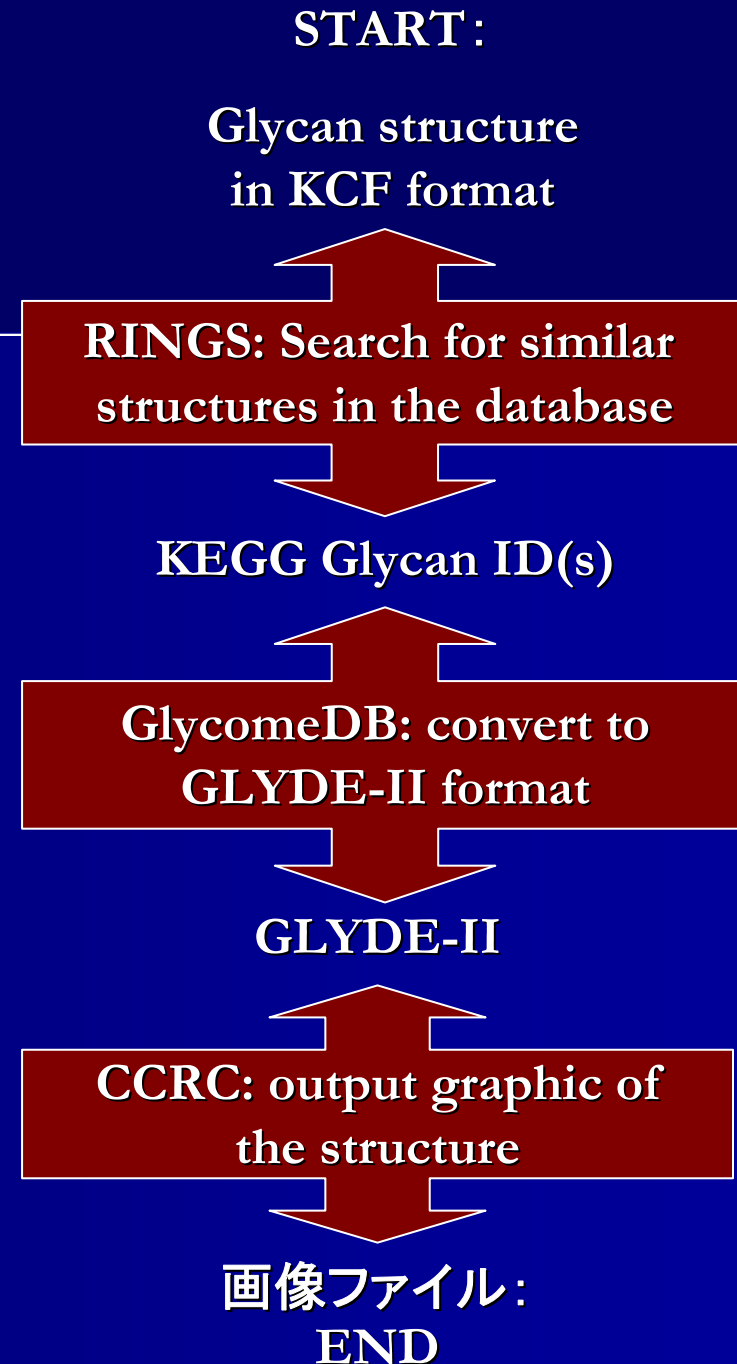
Workflows

- If data types are defined consistently, then different web services can be linked together in order to perform more complex tasks automatically



Actual glyco-workflow

- Toy example of an actual workflow (right)
- Demonstrates the possibilities of workflows for glycobiology





Potential workflows

- Given a particular glycan structure, perform a similarity search for other similar glycans (RINGS) and retrieve links to all other known databases containing the similar structures (GlycomeDB)
- Given MS data for a particular glycan, compute its profile and search for similar profiles in the database (RINGS)
- Many more...



Vision of glycome analysis

CAZY ~ CARBOHYDRATE-ACTIVE ENZYMES

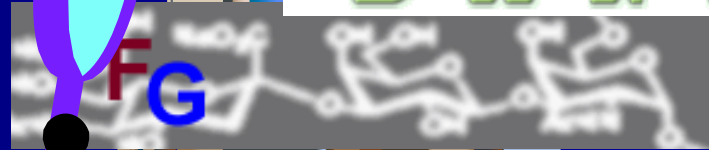
All resources providing web services can be linked together into custom-defined workflows which are available by the click of the mouse.

LECTINES

GOOB

LIPID BANK
for Met

Glycome-



SCIENTIFICS.DE

Leading to discoveries much more efficiently than was possible before.

RESOURCE
INFORMATICS OF
GLYCOMES AT
SOKA

KEGG
GLYCAN

