

Mining Significant Patterns from Trees

Hiroshi Mamitsuka
Bioinformatics Center
Kyoto University

Outline

- Introduction
 - Glycobiology: A specific tree application
- Frequent pattern mining-based approach
 1. Mining alpha-closed frequent subtrees
 - Brief algorithm overview
 - Empirical results
 2. Re-ranking frequent subtrees by hypothesis testing
 - Fisher's exact test
- Summary

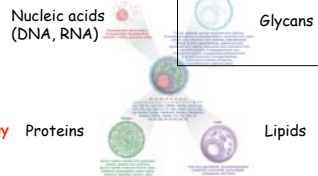
Carbohydrate Chains or Glycans



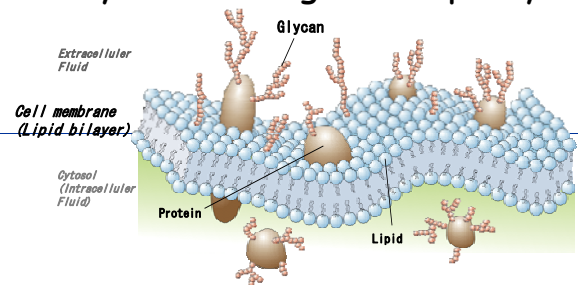
"As indivisible units of life, the cells of all organisms consist of four fundamental macromolecular components: nucleic acids (including DNA and RNA), proteins, lipids, and **glycans**."

J.D.Marth, A unified vision of the building blocks of life, *Nature Cell Biology*, 10(9), Sep. 2008.

Science special issue:
Carbohydrates and Glycobiology
(*Science*, 291(5512), 2001)



Glycans: Biological Property

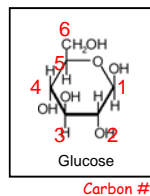


- Usually found on cell surfaces, connecting to glycoproteins
- Crucial to the development and functioning of multicellular organisms

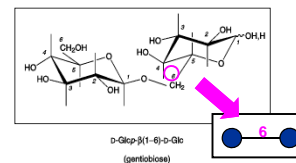
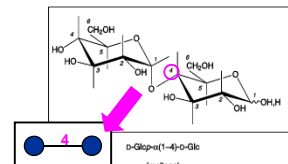
What Does a Glycan Look Like?

- Building blocks: monosaccharides (sugars)

- Galp Galactose
- GalpNAc N-acetylgalactosamine
- Glcp Glucose
- GlcpNAc N-acetylglucosamine
- Manp Mannose
- Fucp Fucose
- Xylp Xylose
- ...



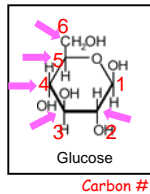
Disaccharides



- Edge, linking two monosaccharides, labeled by **carbon numbers**

What Does a Glycan Look Like?

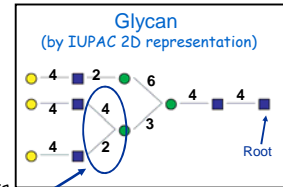
- Monosaccharide connected each other
- No looped connections
- Only one monosaccharide connects to a glycoprotein
 - Resulting in **tree** structures!



What Does a Glycan Look Like (in Summary)?

• Labeled ordered tree

1. Rooted tree
 - Only one monosaccharide connects to a glycoprotein
2. Labeled tree
 - Nodes labeled by monosaccharides
3. Ordered tree
 - Edges labeled by carbon #s and ordered



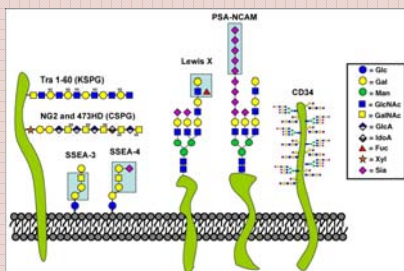
Challenges in Glycobiology

- Complex structure and biosynthesis process
 - Large range of biological functions, from unimportant to crucial for organism survival
 - The same glycan may have different roles
 - Experimentally different results based on different environments (i.e., in vitro or in vivo, etc.)
- ↓
- Informatics must help glycobiology

Glycoinformatics

- Issues:
 - Data collection and database generation
 - Data modeling and glycan structure representation
 - Structure comparison
 - - Mining and prediction algorithms etc.

Conserved Structures in Glycans



Lancot, P.M. et al. (2007) *Curr. Opin. Chem. Biol.*, 11, 373-380

➡ Various substructures known as functional motifs

The Objective!

- Efficiently finding conserved patterns from glycan structures!
- An idea:
 - Frequent pattern mining-based approach

Frequent Pattern Mining-based Approach: Two Steps

1. Mining frequent subtrees
2. Finding significant subtrees from frequent subtrees by hypothesis testing

Frequent Subtrees (1/2)

Definition: Frequent subtree T

$$\text{support}(T) \geq \text{minsup}$$




$\left\{ \begin{array}{l} \text{support}(T): \# \text{trees containing } T \\ \text{minsup} : \text{a threshold for support} \end{array} \right.$

Related work:

- Mining frequent trees
Zaki, M. J. (2002) In *KDD*, 71-80, Edmonton, Canada
- *CMTreeMiner*
Chi, Y. et al. (2005) *IEEE TKDE*, 17, 190-202

Frequent Subtrees (2/2)

- Frequent subtrees obtained from KEGG Glycan database

Support	Subtrees
1365	
1346	
1310	
...	...

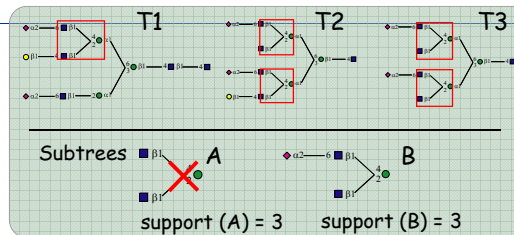
➡ Too many redundant subtrees

Closed Frequent Subtrees (1/2)

Definition: Closed frequent subtree T




$$\text{support}(T) \geq \text{minsup} \ \&$$

$$\text{support}(T) < \text{support}(T'), \ T': \text{a supertree of } T$$



Closed Frequent Subtrees (2/2)

- Closed frequent subtrees obtained from KEGG Glycan database

Support	Subtrees
1365	
1346	
1310	
...	...

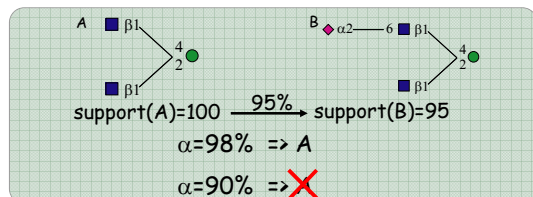
➡ Still too many redundant subtrees

Proposed Idea: α -closed Frequent Subtrees [ECCB08]

Definition: α -closed frequent subtree T

$$\text{support}(T) \geq \text{minsup} \ \&$$

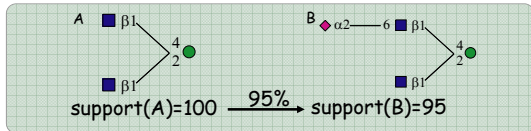
$$\text{support}(T) < \max(\alpha \times \text{support}(T'), \text{minsup})$$



Nice properties of α -closed Frequent Subtrees

If $\alpha = 1$

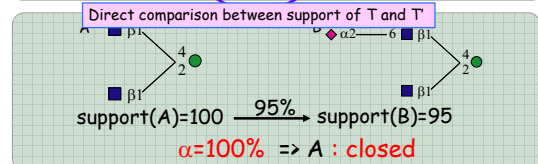
Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(\alpha \times \text{support}(T), \text{minsup})$



Nice properties of α -closed Frequent Subtrees

If $\alpha = 1$

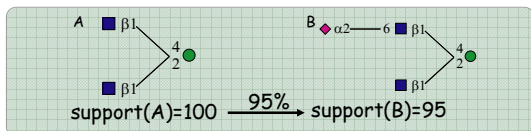
Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(1 \times \text{support}(T), \text{minsup})$



Nice properties of α -closed Frequent Subtrees

If $\alpha = 0$

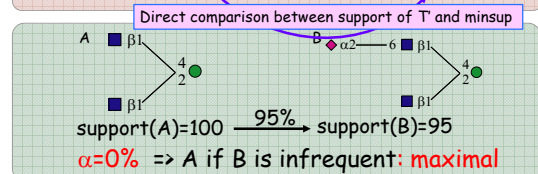
Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(0 \times \text{support}(T), \text{minsup})$



Nice properties of α -closed Frequent Subtrees

If $\alpha = 0$

Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(0 \times \text{support}(T), \text{minsup})$

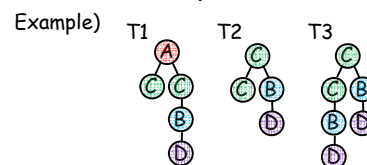


Summary: Nice properties of α -closed Frequent Subtrees

Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(\alpha \times \text{support}(T), \text{minsup})$

- If $\alpha=1$, α -closed frequent subtrees are closed frequent subtrees.
- If $\alpha=0$, α -closed frequent subtrees are maximal frequent subtrees.
- α -closed frequent subtrees are a monotone increasing family from 0 to 1.

Enumeration of Frequent Subtrees



minsup = 2
 \Rightarrow Subtree T is frequent if $\text{support}(T) \geq 2$, where
 $\text{support}(T) = \#\text{trees including at least one T}$
 eg. $\text{support}(\begin{matrix} \text{C} \\ | \\ \text{C} \end{matrix}) = 3$

Use an enumeration tree for efficiency

Enumeration of Frequent Subtrees

Subtrees with one node:

A	B	C	D
support = 1	support = 3	support = 3	support = 3
(T1 only)	(T1, T2, T3)		

Downward closure property!
patterns including an infrequent subtree are always **infrequent**

(no need to check)

minsup = 2

Enumeration of Frequent Subtrees

Grow frequent subtrees only

minsup = 2

Enumeration of Frequent Subtrees

Grow frequent subtrees further

minsup = 2

Enumeration of Frequent Subtrees

Grow frequent subtrees further

minsup = 2

Enumeration of Frequent Subtrees

Complete frequent subtree set

B	C	D
support=3	support=3	support=3
B-C	C-B	C-C
support=3	support=3	support=2
B-C-B	C-B-C	C-C-B
support=3	support=2	support=2
B-C-D	C-B-D	C-C-D
support=3	support=2	support=2

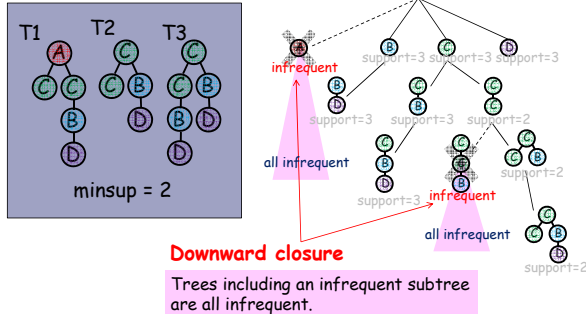
minsup = 2

Enumeration of Frequent Subtrees

Complete frequent subtree set over the enumeration tree

minsup = 2

Pruning Enumeration Tree 1. Downward Closure



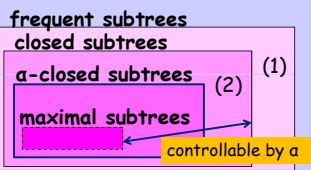
Nice properties of α -closed Frequent Subtrees

Definition: α -closed frequent subtree T
 $\text{support}(T) \geq \text{minsup}$ &
 $\text{support}(T') < \max(\alpha \times \text{support}(T), \text{minsup})$

- If $\alpha=1$, α -closed frequent subtrees are closed frequent subtrees.
- If $\alpha=0$, α -closed frequent subtrees are maximal frequent subtrees.
- α -closed frequent subtrees are a **monotone increasing family from 0 to 1.**

Pruning Enumeration Tree

Frequent subtree hierarchy



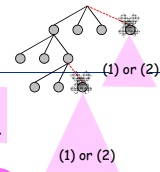
Pruned by downward closure property

- We can further prune:
- (1) frequent but not closed
 - (2) closed but not α -closed

Pruning Enumeration Tree 2. Closedness and α -closedness

We can prune:

- (1) frequent but not closed
- (2) closed but not α -closed

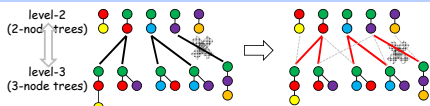


We can prune the branch if all subtrees are (1) or (2).

- (1) Check closedness
- (2) Check α -closedness

Pruning by Closedness or α -closedness

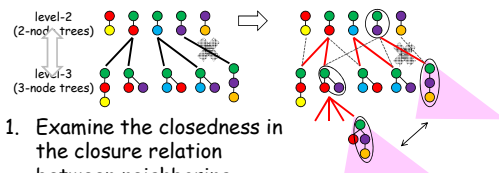
Compare frequent subtrees between neighboring levels



- (1) Closedness
Use Left- and Right-blanket pruning (Chi et al., 2005) of closed frequent subtrees
- (2) α -closedness
Unless α -closed, there can be frequent supertree T' with " $\text{support}(T') \geq \alpha \times \text{support}(T)$ " in the next level.

Pruning by Closedness: A Bit Detail

Compare frequent subtrees between neighboring levels



1. Examine the closedness in the closure relation between neighboring levels.
2. If found, a smaller subtree can be discarded.

Frequent ≠ significant

- Frequent subtrees obtained from KEGG Glycan database

Support	Subtrees
1646	● β1 — 6 ●
1628	● β1 — 4 ■
1365	■ β1 — 4 ■
⋮	⋮

➔ Significance test against control needed!

Statistical Hypothesis Testing (1: Procedure)

1. Generate synthetic control (negative)
 - For each frequent pattern, count #appearances in both datasets
2. Apply Fisher's exact test to compute P-value
3. Rank by P-values

Statistical Hypothesis Testing (2: Contingency Table)

2x2 contingency table for frequent subtree T

	Positives	Controls	Total
With T	# true positives (n_{TP})	# false positives (n_{FP})	n_T
Without T	# false negatives (n_{FN})	# true negatives (n_{TN})	$n_{\bar{T}}$
Total	n_P	n_N	n

Statistical Hypothesis Testing (3: Fisher's Exact Test)

Probability from this table in random fashion:

$$Pr = \frac{\binom{n_T}{n_{TP}} \binom{n_{\bar{T}}}{n_{FN}}}{\binom{n}{n_P}}$$

$$= \frac{n_T! n_{\bar{T}}! n_P! n_N!}{n! n_{TP}! n_{FP}! n_{FN}! n_{TN}!}$$

follows hypergeometric distribution.

➔ p -value can be computed by using the CDF of hypergeometric distribution!

Frequent pattern mining-based approach

Result 1

Significant patterns examined

Data and Parameter Setting

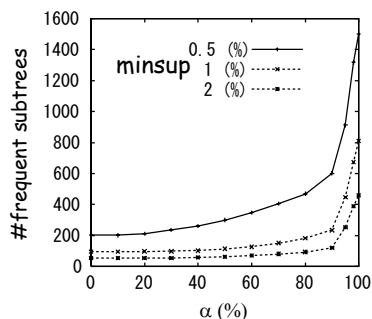
1. Data

- > 7454 positives from KEGG GLYCAN
- > 7454 synthetic control, being kept the same as positives in
 - topology
 - parent-child pairs

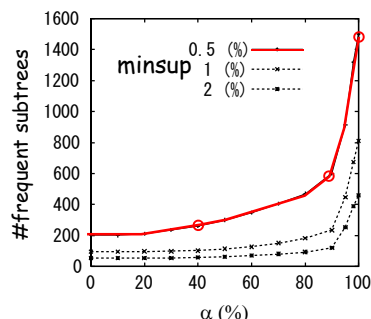
2. Parameter settings

- > alpha (0~100%)
- > minsup (0.5%, 1%, 2%)

#frequent Subtrees versus α



#frequent Subtrees versus α



Top 5 Significant Subtrees Ranked by P-values

Rank	P-value	Support	Substructures	Name
1	1.6e-46	381		Lewis X
2	1.1e-40	164		O-glycan core
3	5.0e-26	109		Glycosphingolipid core lactoseries
4	5.6e-26	233		Glycosphingolipid core lactoseries
5	8.2e-26	83		Lewis A

alpha=40%, minsup=0.5%

Top 6-10 Significant Subtrees Ranked by P-values

Rank	P-value	Support	Substructures	Name
6	1.3e-24	79		N-glycan core high mannose type
7	2.7e-24	78		N-glycan core complex type
8	2.9e-21	68		N-glycan core complex type
9	2.9e-21	68		N-glycan core complex type
10	3.2e-20	74		Blood group H

Frequent pattern mining-based approach

Result 2

Classification performance examined

Applying Significant Patterns to Classification for Comparison

Problem setting:

> Discriminate real O-glycans (positives) from randomly synthesized, almost similar trees (negatives or controls).

Input: $T = (0, 1, 1, 0, 1, \dots, 1, 0)'$

(i-th element is 1 if T has i-th significant subtree)
0 otherwise

Note: input vector size controllable by $d!$

Data and Procedure

Data

- 485 positives: O-glycan structures
- 485 synthetic control

Procedure

- linear SVM
- 10-fold cross-validation being averaged over 10 random controls
- Use α -closed patterns obtained from the training data

Competing Methods

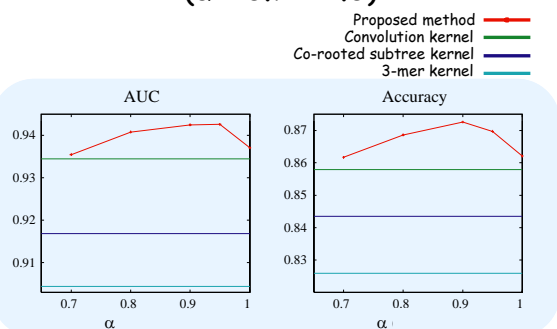
SVM with tree kernels

- ➔ Well-recognized as high-performance method in the current machine learning

Three cutting-edge tree kernels

- Convolution kernel
(Kashima, H. *et al* 2002)
- Co-rooted subtree kernel
(Shawe-Taylor, J. *et al* 2004)
- 3-mer kernel
(Hizukuri, Y. *et al* 2005)

Performance Results ($\alpha = 0.7 \sim 1.0$)



AUC and Accuracy

Method	AUC (<i>P</i> -value)	Accuracy (<i>P</i> -value)
Proposed method	0.942	0.869
Convolution kernel	0.934 (6.91e-03)	0.857 (1.14e-02)
Co-rooted subtree kernel	0.916 (7.78e-11)	0.843 (1.03e-06)
3-mer kernel	0.904 (4.74e-18)	0.825 (9.91e-15)

P-value ($\alpha = 95\%$) in parenthesis

Outperformed all competing methods, being statistically significant!

Summary

- The new concept of α -closed frequent subtrees proposed and efficient algorithm for mining α -closed frequent subtrees presented
- Mining frequent subtrees combined with hypothesis testing for finding significant subtrees
- Performance confirmed experimentally
 - Existing significant patterns detected correctly
 - Outperformed competing methods, including SVM with tree kernels
- Further analysis on found patterns ongoing

Acknowledgements

Kyoto University
Bioinformatics Center
Institute for Chemical Research

- Kosuke Hashimoto
- Ichigaku Takigawa
- Motoki Shiga
- Minoru Kanehisa

Reference Information

- [ECCB08]
 - Mining Significant Tree Patterns in Carbohydrate Sugar Chains. *Bioinformatics*, 24 (16) (Proceedings of the Seventh European Conference on Computational Biology (ECCB 2008), Cagliari, Sardinia-Italy, September, 2008), i167-i173, 2008.

Funding Acknowledgements

- BIRD, JST
- Research fellowship for Young Scientists, JSPS
- Grant-in-Aid for Young Scientists, JSPS
- Kyoto University 21st Century COE program, Knowledge Information Infrastructure for Genome Science, JSPS
- Education and Research Organization for Genome Information Science, JST