Analysis of Glycan Data using Non-negative matrix factorization

Ryo Hayase, *Graduate School of Science and Technology, Keio University*

Background

Glycans are the compounds which are formed by sugars linking a chain. They are crucial for many key biological processes and their alterations are often a hallmark of disease.

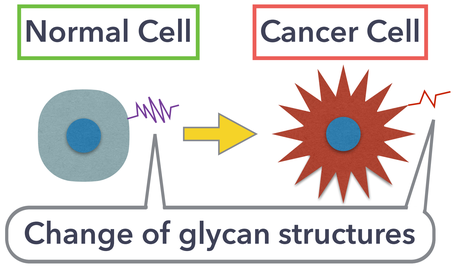

Fig.1 Normal/Cancer cell's surface

Therefore glycans is used as tumor marker.
What is the reason why new tumor marker is necessary?
There are two reasons for the above:
①検査を簡単に行えるようにするため
②false positive rateやfalse negative rateを減らすため
すでに腫瘍マーカーは数多く報告されているが新しいものを探す理由がここにある

Objectives

The purpose of this research is to discover the glycans that could be the tumor markers candidate.
We analyze the glycan data using Non-negative matrix factorization to search tumor markers candidate in order to be easily understand a relationship of glycans and cancers.

Table.1 The glycan data (provided by Prof.Oya and Prof.Sato)

	Cancer cell	Gastric			Prostate				Lung		Liver	Breast	
	Structure (P:GalNAc-Thr-C12)	HuG1-N	NUGC4	MKN45	LNCaP	C4-2	KO18	AT18	A549	HLC1	Huh-7	MDA-MB-453	SK-BR-3
1	NeuAc-P	+++	++	+++	+	+			+	+	++	++	+++
2	Hex-P	+++	+++	+++	+++	+++	+++	+++	++	+	++	+++	++
3	HSO$_3$-Hex-P	++	+++						+		++	++	
4	Fuc-Hex-P	+++	++	++	++	+	+++	++					
5	HexNAc-Hex-P			+									
6	HexNAc-（Fuc-)Hex-P	++	+	++									
7	Hex-HexNAc-Hex-P			++									
8	Hex-(Fuc-)HexNAc-Hex-P	+	++	+									
9	NeuAca2-3Hex-P	++	++	++	++	++	+++	++	+	+	+++	+++	+++
10	Hex-(NeuAca2-6)P	++	++	++									++
11	NeuAc-(HSO$_3$-)Hex-P	+	+						+			++	
12	Fuc-Hex-(NeuAca2-6)P	++		+									
13	NeuAc-Hex-HexNAc-Hex-P			+									
14	NeuAc-Hex-(NeuAc-)P	+++	+++	++	++	+	++	++	+	+	+++		+++
15	NeuAc-Hex-(Fuc-)HexNAc-Hex-P			++									
16	NeuAc-Hex-HexNAc-Hex-(NeuAc-)P	++		++									
17	NeuAc-Hex-(Fuc-)HexNAc-Hex-(NeuAc-)P	+++		++									
18	Hex-(HexNAc-)P	++	+	++	++	++	+++	++					
19	Fuc-Hex-(HexNAc-)P	++	+	+	+		++	++					
20	Hex-(Hex-HexNAc-)P	+		+	++	++	+++	++					
21	Fuc-Hex-(Hex-HexNAc-)P	++											
22	Hex-(Hex-（Fuc-)HexNAc-)P		+	+	+	+	+++	++					
23	NeuAc-Hex-(HexNAc-)P	++	+	+	+		+				++	++	++
24	NeuAc-Hex-(NeuAc-HexNAc-)P										+++		
25	NeuAc-Hex-(Hex-HexNAc-)P	+		+	++	+	+++	++					
26	NeuAc-Hex-(Fuc-)HexNAc-Hex-)P	++	+				+++	++					
27	NeuAc-Hex-(NeuAc-Hex-HexNAc-)P	+		++	+		+	+	+	+	++		
28	NeuAc-Hex-(Fuc-)HexNAc-(NeuAc-Hex-)P						+		+				
29	HexNAc-P	++	++	++	+++	++	++	+++			++	+	+++
30	HexNAc-(HexNAc-)P		++		++	+	++	++					
31	Hex-HexNAc-(HexNAc-)P				++	+	+++	++					
32	Fuc+Hex+2HexNAc+P				+		++	++					

immeasurably small
+ Area/mg Protein $< 1 \times 10^7$
++ $1 \times 10^7 \leq$ Area/mg Protein $< 1 \times 10^8$
+++ $1 \times 10^8 \leq$ Area/mg Protein

Columns represent the 12 types of cancer cells of 5 cancers. Rows represent 32 different types of glycans. Original data are numerical values.

Purpose : Searching for tumor markers candidate

Methods

Non-negative matrix factorization (NMF) was first proposed by Lee and Seung (1999).
NMF is an algorithm that is used decompose data matrix V into basis matrix W and coefficient matrix H ,with the property that three matrices have non-negative elements.
It can be used methods of feature quantity extracting.
Consider the NMF model, given by:

NMF model

$$\underset{n\times m}{V} \approx \underset{n\times r}{W} \times \underset{r\times m}{H}$$

data matrix $\in \mathbb{R}_+^{n\times m}$; *basis matrix* $\in \mathbb{R}_+^{n\times r}$; *coefficient matrix* $\in \mathbb{R}_+^{r\times m}$ $(r \leq n, m)$

$\mathbb{R}_+^{n\times m}$ represents a space of n by m matrices with non-negative elements.
Where basis number r is smaller than n and m.
So W and H are smaller than V in the sense of matrix size.

NMF problem

$$\underset{W,H}{\text{minimize}} \parallel V - WH \parallel_F^2$$
$$\text{subject to } W, H \text{ is non-negative}$$

In this research, we used the Frobenius norm to measure the distance between V and WH.
We applied "Hierarchical Alternating Least Square"(HALS) algorithm to solve this problem.
What kind of knowledge will be provided from the decomposition of NMF?

basis matrixおよびcoefficient matrixは分解前のデータ行列の行および列の情報を持っている。

At the glycan data, basis matrix has information about the glycan, coefficient matrix has information about the cancer.
So we can obtain each information by analysis for W and H.

$$\underset{32\times 12}{V} \approx \underset{32\times 5}{W} \times \underset{5\times 12}{H}$$

data matrix ; *basis matrix* ; *coefficient matrix*

Information about the… Glycan Cancer

Analysis of the glycan data

We set a basis number to 5 (same as number of cancers).
As mentioned above, basis matrix has information about the glycan, coefficient matrix has information about the cancer.
Fig.2(A) is the heatmap of the coefficient matrix. We observed that, gastric cancer is characterized by basis 1 and 2, prostate cancer is characterized by basis 3, lung cancer is characterized by basis 4, liver cancer is characterized by basis 4 and 5, breast cancer is characterized by basis 5.
In fact, clustering of each column of coefficient matrix classified cancers well. (Fig.3)
Fig.2(B) is the heat map of the basis matrix. Finally, we searched for tumor markers candidate from basis matrix.
We interpreted basis matrix in Fig.2(B). ★ are existing tumor markers. So we searched tumor markers candidate other than these.
特徴的な基底(basis)が各がんにあった。そこで各糖鎖(glycans)の値が大きいもの(0.15以上のもの)をその基底に対し特徴があると考え、各がんに対して腫瘍マーカーとなりそうな糖鎖を探した。
As a result, glycans of ★ were chosen. (Fig.4) Possibly these glycans may be used as tumor marker in the future.

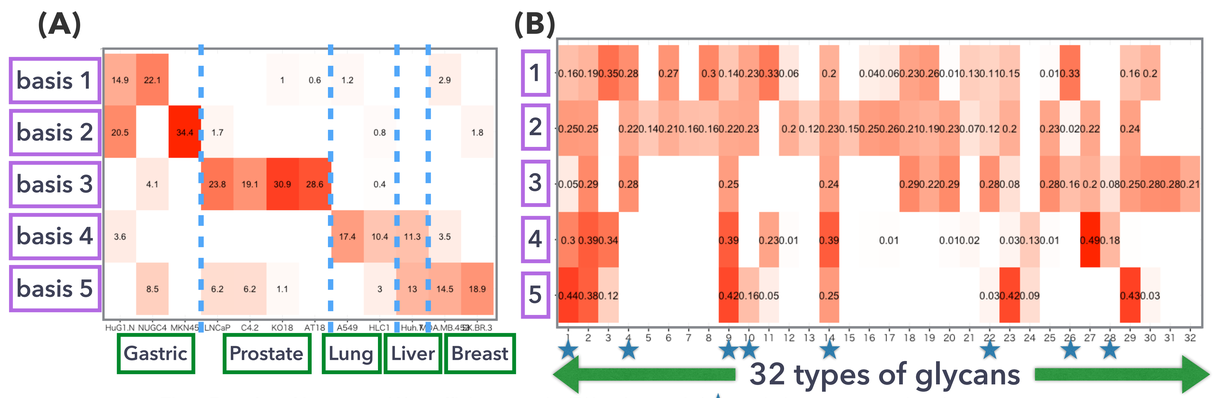

Fig.2 Results of heatmap (A)coefficient matrix (B)basis matrix(★: existing tumor markers)

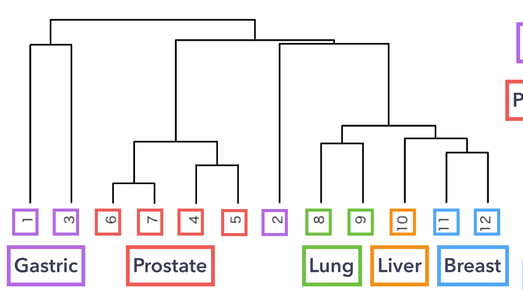

Fig.3 Clustering of coefficient matrix
(Using "single-linkage" methods)

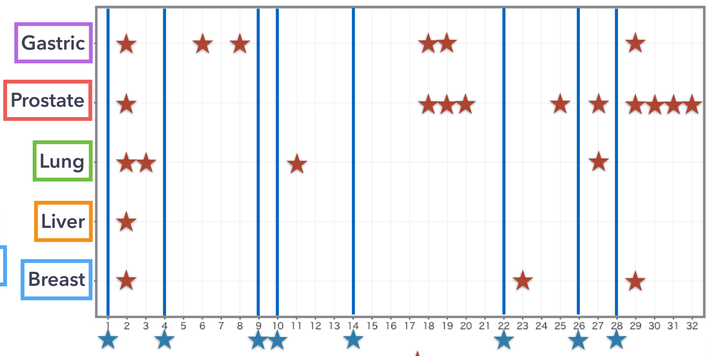

Fig.4 Interpretation of basis matrix(★: tumor markers candidate)

Conclusion

As a result of the factorization of the glycan data…

- From a coefficient matrix, we were able to classify cancers well.
- From a basis matrix, we were able to search the glycan which is the tumor marker candidate.

Acknowledgement

We thank Prof. Sato (Department of Bioscience and Informatics, Keio University) and Prof. Oya (School of Medicine, Keio University) for providing the glycan data.

References

[1] Lee DD and Seung HS. (1999). Learning the parts of objects with nonnegative matrix factorization. *Nature*. vol. 401, pp. 788-791.
[2] Lee DD and Seung HS. (2000). Algorithms for Non-negative Matrix Factorization. *In NIPS*. Vol. 13, pp. 556-562.
[3] A. Cichocki and A. Phan. (2009). Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. *IEICE Trans,* Vol.E92-A, No.3, pp.708-721