Constrained clustering for discovering high binding affinity glycan substructures

Haruna Terai1  Jun Sese1
terai@sel.is.ocha.ac.jp  sesejun@sel.is.ocha.ac.jp

1 Dept. of Computer Science, Ochanomizu University, 2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Keywords: clustering, glycan

1 Introduction

While information of the gene has increased by the advancement of DNA and the protein analysis, glycans are also known to exert a big influence on the vital activity. Because the slight differences of their structures greatly influence the binding affinities with the proteins and viruses, it is important to know the structures related to the bindings. In this research, we proposed a method to estimate the substructures recognized when proteins or viruses binds to glycans. Recently, it is possible to measure the binding affinity of 377 kinds of carbohydrate structures by glycan array at the same time. By using this data, we developed a clustering method constrained by the carbohydrate chain substructures to discover the combination of substructures that relates to binding which is adjacent or not. Proteins are known to bind to the end of the carbohydrate chains, but there are secondary sites, which have been found to be important determinants of lectin binding specificity [3]. In this research, we developed a clustering method under the constraint of carbohydrate structures. By using the method, we can discover plural binding sites shared by the series of experimentations.

2 Method and Results

2.1 Method

In this section, we will define the index to extract the substructure that shows a high binding affinity by structure specificity.

Glycan array data. Like the microarray used for gene expression analysis, there is glycan array that can measure the binding affinity with the protein and the virus of all glycans on the slide at a time by making about 300 glycan spots on the slide glass. The result data consists of glycan number, glycan name and binding affinity. This data is open to the public in Consortium for Functional Glycomics(CFG) [2, 5].

Constrained clustering. By using the glycan array, we will examine which substructure influences the binding affinity of each experiment. To extract the high binding affinity glycan structure, we will define a index to set the standard of the binding affinity. Since glycan array is an experiment constructed outside the cell, there is a possibility of binding or a noise not seen in the cell. To consider these possibilities and to extract an appropriate partial structure, we will set the next three conditions. The binding affinity level is large, the difference of the binding affinity level is low under the same experiment, and the binding affinity level changes as the density of reagent changes. The first condition is appreciable by calculating the average of the binding affinity level that has the substructure S, this is assumed to be $\mu(S)$. The second condition is appreciable by calculating the variance. To make
them correspond to plural experiments, the variance is extended to the average of the variance. This is assumed to be $\sigma(S)$. The last condition is appreciable by calculating the difference of the average of the binding affinity level in each experiment, this is assumed to be $\delta(S)$.

To choose the substructure to meet all three requirements, the goodness of a structure $S$ is measured by the following index $gindex(S)$.

$$gindex(S) = \frac{\mu(S) \delta(S)}{\sigma(S)}$$

In this index, the substructure that has large overall average $\mu(S)$ and change amount $\delta(S)$, and small variance $\sigma(S)$ will have a bigger value. By calculating the $g$-index for all substructures, we can rank the substructures which specificity binds to proteins and viruses. Furthermore, we took the union of substructures to discover plural binding sites shared by the series of experiments.

2.2 Results

We analyzed a glycan binding protein Galectin-1 and a variant of Influenza B virus. Galectin-1 is known to bind the poly-N-lactosaminyl glycan structures $(3\text{Gal}\beta1,4\text{GlcNAc}\beta)n$ [3]. Influenza B virus is known to bind the N-acetyleneuraminic acid structure Neu5Acα- and especially known that influenza B virus strongly recognize the structure consisting $(\text{Neu5Ac}2-6(3)\text{Gal}\beta1-3(4)\text{GlcNAc}\beta1-) [4]$. We confirmed that our method estimated these substructures.

3 Conclusion

In this research, we analyzed the substructures which are recognized when the protein and the virus binds to the glycans. We paid attention to the presence of the substructure of the glycan and developed a index that can discover a substructure that rises the binding affinity in two or more experiments. Finally, we confirmed that our method could estimate the known carbohydrate structures recognizing Galectin-1 and Influenza B virus. Moreover, there might be two or more substructures recognized when the protein binds with the sugar chain, and it is possible to happen to not only the end but also the middle part of the sugar chain.

References


