

# Prediction of the *O*-glycosylation and Environments around the crowded and isolated *O*-glycosylation sites

Yukiko Nakajima<sup>1</sup>                      Hirotaka Sakamoto<sup>1</sup>                      Kazutoshi Sakakibara<sup>1</sup>  
nakajima@sys.ci.ritsumei.ac.jp      sakamoto@sys.ci.ritsumei.ac.jp      sakaki@sys.ci.ritsumei.ac.jp

Masahiro Ito<sup>2</sup>                      Ikuko Nishikawa<sup>1</sup>  
maito@sk.ritsumei.ac.jp      nishi@ci.ritsumei.ac.jp

<sup>1</sup> College of Information Science and Engineering,

<sup>2</sup> College of Life Science, Ritsumeikan University, Kusatsu, Shiga 525-8577 Japan

**Keywords:** *O*-glycosylation, prediction, support vector machine, motif, crowded and isolated sites

## 1 Introduction

Glycans play an important role in various life phenomena by combining to protein or lipid. *O*-glycosylation is one of the main two types of the mammalian protein glycosylation. It is serine (Ser) or threonine (Thr) specific, though any consensus sequence is still unknown, while the binding process and a consensus sequence is clarified for *N*-glycosylation. We have been applied support vector machines (SVM) for the prediction of *O*-glycosylation sites from various kinds of protein information, aiming to investigate a condition of the *O*-glycosylation and elucidate the mechanisms[1]. In the present study, we focus on the distribution of the glycosylation sites. It is known that there are crowded and isolated glycosylation sites, and these two types are considered to be generated by different mechanisms. Wilson et al.[2] compared the protein sequences around these glycosylation sites. In this report, 98 mammalian protein sequences are obtained from UniProt12.2, and SVM is trained to predict the crowded and isolated *O*-glycosylation sites separately, to find any difference.

## 2 Method and Results

### 2.1 Prediction by SVM

98 mammalian protein sequences are selected from UniProt12.2, which contain some Ser and Thr residues annotated as being *O*-glycosylated experimentally. There are 452 Ser or Thr sites with annotation, together with 6004 sites without annotation, which are denoted positive and negative, respectively. Many positive sites are found densely crowded, while others are found sparsely. We define two types of positive sites as follows. That is, if the nearest neighbor Ser or Thr site in either side is glycosylated, then it is called crowded glycosylation site. Otherwise, it is called isolated. Then among 452 positive sites, 307 sites are found crowded, while the rest 145 sites are isolated. SVM is trained for each type separately.

Protein primary sequence with length  $W_s$  is used for the input data to SVM, with Ser or Thr site at the center as a prediction target. Each residue is encoded by 21 bit-sparse coding, which distinguishes all 20 kinds of amino acid and a null outside a terminal. RBF kernel is used for SVM, and  $W_s$  are varied from 3 to 55. The above number of positive sites and similar number of negative sites are segmented into 10 sets for 10 fold cross-validation.

Fig.1 shows the numbers of correctly and falsely predicted positive sites, both for isolated and crowded groups. The prediction efficiency increases according to  $W_s$  with the highest value at  $W_s = 45$

for the crowded sites. On the other hand, the correct ratio stays rather constant for the isolated sites, including at  $W_s = 3$ . Therefore, the detailed primary sequences around the positive sites are compared between the two types in the followings.

## 2.2 Primary Sequences around isolated and crowded glycosylation sites

Fig.2 shows the existence ratio of proline (Pro) at each position within  $W_s = 31$  around the crowded positive, the isolated positive and the negative sites. Pro has been well known[2] to have a high

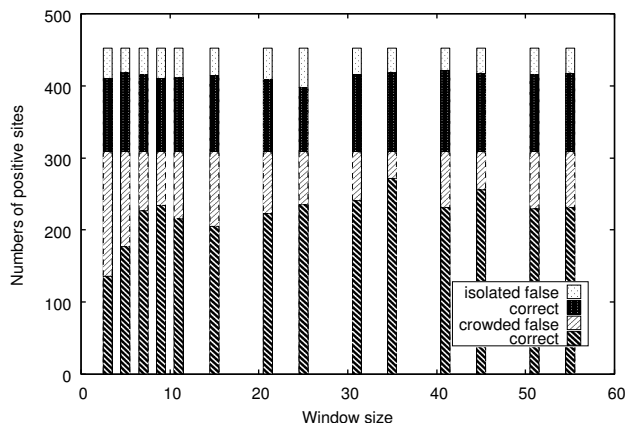


Figure 1: Numbers of correctly and falsely predicted positive sites for the isolated and crowded groups

probability at  $-1$  and  $+3$ . Fig.2 shows the high values at  $-1$  especially for the isolated positive, and at  $+3$  for both types of the positives. This leads to the high correctness even at  $W_s = 3$  for the isolated positive seen in Fig.1.

Similarly for valine (Val), the very sharp peaks of 16% both at  $-3$  and  $+8$  are observed only for the isolated positive. Other peaks are found also for alanine (Ala) at  $-6$  and  $+5$ , especially for the isolated positive. Thus, some motifs are expected for the isolated positive.

On the contrary, the interaction between the glycosylated sites and a gross amino acid composition over the range up to  $W_s = 31$  is expected to affect the glycosylation for the crowded positive, which is now investigated in more detail. Both positive types are found mainly in coil structure, other than helix or sheet, which implies the glycosylation contributes to the stabilization of the protein structure.

## 3 Discussions and Future Works

More details are under investigation for any motif around the isolated positive site and any relevant composition around the crowded positive site to elucidate the mechanisms for each type. The future works include the prediction of an unknown positive site followed by biological experiments.

## References

- [1] Sakamoto, H. et al., Prediction of Mucin-type *O*-glycosylation using Structure Information by Support Vector Machines, *Genome Informatics*, 18:55–56, 2007.
- [2] Wilson I. et al., Amino acid distributions around *O*-linked glycosylation sites, *Biochem. J*, vol.275, pp.529–534, 1991.

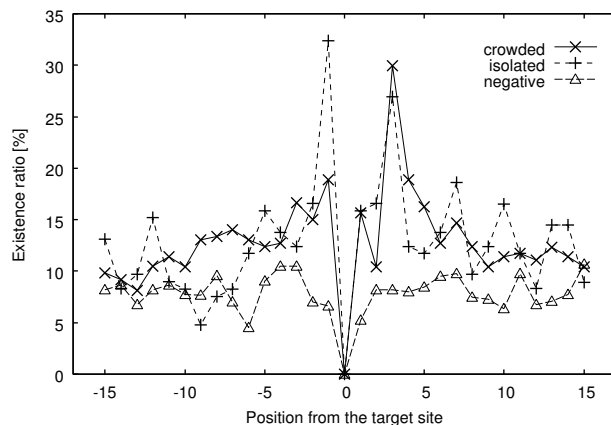


Figure 2: Existence ratios of Pro at each position around crowded positive, isolated positive and negative site