Predicting Human Protein Phosphorylation Sites in Intrinsically Disordered Region by Support Vector Machine

Ikuko Nishikawa[1§], Tomoki Ishino[1], Yukako Tohsato[2], Shuichi Onami[2,3], Satoshi Fukuchi[4] and Ken Nishikawa[5]

[1] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan
[2] RIKEN Quantitative Biology Center, Kobe, Hyogo, Japan
[3] National Bioscience Database Center, Japan Science and Technology Agency, Japan
[4] Department of Bioinformatics, Maebashi Institute of Technology, Maebashi, Gunma, Japan
[5] Institute for Protein Research, Osaka University, Suita, Osaka, Japan
[§]Corresponding author nishi@ci.ritsumei.ac.jp

Key words: human protein phosphorylation, intrinsically disordered region, support vector machine, evolutionary conservation

1. Introduction

Phosphorylation is one of the most important and widespread post-translational modifications, and revealing its mechanism is important for understanding the life phenomena. In this study, phosphorylation sites in human proteins are predicted by support vector machine (SVM). Among several machine learning approaches for the prediction of phosphorylation sites (Blom et al., 2004), the present research focuses on the phosphorylation in intrinsically disordered region (IDR). Then, it is revealed that simple sequence information is enough for the effective prediction of the phosphorylation sites in the functional domain, while it is not the case in IDR. This is reasonable when we consider a relatively low sequence conservation property of IDR. Then, the evolutionary conservation of each amino acid residue in IDR is examined, especially for phosphorylated and non-phosphorylated Ser/Thr. Phosphorylation sites are further classified into with and without any experimentally observed function.

2. Method and Results

2.1 Protein Data

Phosphorylation data and amino acid sequence data of human proteins are obtained from PhosphoELM 9.0. Data of IDR of human proteins is obtained from DICHOT. Input to SVM is an amino acid sequence around a prediction target. Length of input sequence is denoted window size *W*s in the following.

2.2 Prediction of Phosphorylation Sites in IDR/Domain

First, two types of SVM are trained to predict whether each Serine (Ser) or Threonine (Thr) site is phosphorylated, each for phosphorylation sites in IDR region and in the domain. The same numbers of phosphorylated and non-phosphorylated Ser/Thr sites are selected from IDR and domain. Each residue is coded by sparse coding, where 20 kinds of amino acids are coded by 20 bits sequence, plus one more bit for a null. Then, the obtained accuracy is obviously higher for domain. The highest accuracy in domain is 78% at *W*s=35 for Ser, and is 75% at *W*s=31 for Thr. This is comparable with the accuracy obtained by various input information by DISPHOS using neural networks (Iakoucheva et al., 2004). On the contrary, simple sequence information around prediction sites is effective enough for the domain, while it is not for IDR whose accuracy remains around 70%. This may be caused by the relatively low sequence conservation property generally known for IDR, though the phosphorylation is much more abundant in IDR. Therefore, the evolutionary conservation of each amino acid residue in IDR is calculated in the next section.

2.3 Prediction of Functional Phosphorylation Sites in IDR

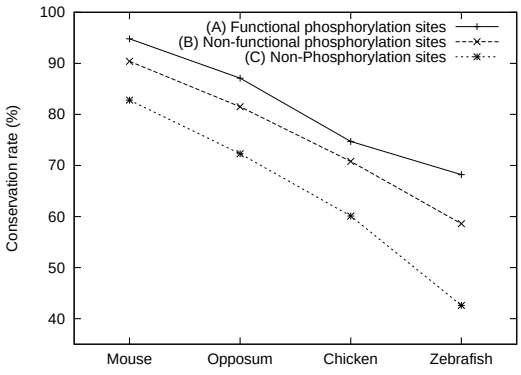
As sequence conservation is generally low in IDR, a target protein and its orthologs are used to give a clear definition of a conservation of each residue. The human (H) phospho-proteins which possess the ortholog in four other vertebrates, that is, mouse (M), opossum (O), chicken (C) and zebrafish (F), are selected. Then, multiple alignment (MA) is obtained by MAFFT for the orthologs, to observe whether each Ser/Thr residue on human protein is conserved in the ortholog for each species. Fig. 1 shows an example human protein Q13562 (NDF1_HUMAN) as a query and its four orthologs after the alignment.

```
H ···SPLTDCTSPSFDGPLSPPLEDDLETMN--AEEDSLRNGGEEED------LAGAQSHGSIF-SGTAAPRCEIPIDNIMSFDSHSHHERVMSAQLNA···
M ···SPLTDCTSPSFDGPLSPPLEDELEAMN--AEEDSLRNGGEEEE------LAGPQSHGSIFSSGAAAPRCEIPIDNIMSFDSHSHHERVMSAQLNA···
O ···STLTDCTSPSFDGPLSPPLDDDLEAMN--PEEESLRNGVEEED------LAGAQGHGSIF-SGTAAPRCEIPIDNIMSYDSHSHHERVMSAQLNA···
C ···GGLPEGAGPAFDGPLSPPLEEDLEALHGEAEEDALRNGEEEDE------LPAAPAHAAVF-SGAAA-RCELPADGLAPYEGHPHHERVLSAQLSA···
F ···TTLTDCTSPSFDGPLSPPL----------TSEPMLNDMEDDDDAGLNRLLAGAQGHAASLYAGSTQ-RCDIPMENIMSYDGHSHHERVMNAQLNA···

     *     * *********           * *             *     *       *    **  *        * *****  *** *
```

Fig. 1 Human protein Q13562 (position 258-352) and four orthologs in mouse, opossum, chicken and zebrafish. Phosphorylated two Ser in IDR are shown in red. Domain is shown in blue.

Ser/Thr in IDR is classified into the following three groups: (A) phosphorylation sites whose function are clearly identified, (B) phosphorylation sites whose function are still not identified, and (C) non-phosphorylation sites. The function of each phosphorylation site is identified from "Description" entry in UniProt. Fig. 2 shows the conservation rate of each group for each species, which is obviously different among (A), (B) and (C). The numbers of residues are 233, 596 and 7287 for (A), (B) and (C), respectively, which are obtained from 133 human phospho-proteins with the four orthologs.



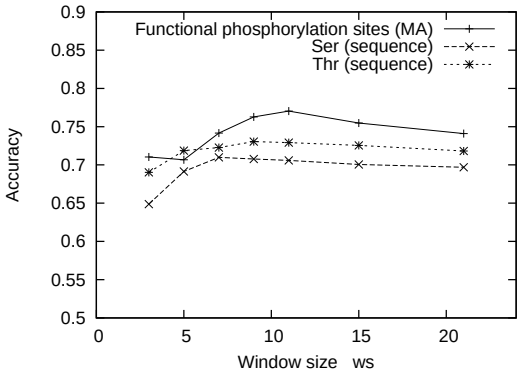


Fig. 2 Conservation rate of Ser/Thr in IDR …..Fig. 3 Accuracy of SVM by multiple alignment sequences

Finally, SVM is developed to classify (A) functional phosphorylation Set/Thr sites and (C) non-phosphorylation Set/Thr sites in IDR. Based on the different conservation rate shown in Fig.2, multiple alignment sequence is used as input to SVM. Fig. 3 shows the improved accuracy of 77% at W_S=9 for Ser/Thr in IDR, compared with the accuracies obtained by single query sequence. This indicates that sequence information of a target human protein is still useful for the prediction, when it is used in company with multiple sequences of the appropriate orthologs, which contain evolutionary information of each site.

3. Discussions and Future Works

The existence of non-functional phosphorylation has been under discussion (Levy et al., 2012), and it still needs the further study to consistently understand the abundance of the phosphorylation in IDR, the low evolutionary conservation in IDR, and the precise regulation of transcription and translation. One of our ongoing studies is to predict the functional sites among the reported phosphorylation sites in IDR.

References

Blom, N., Ponten, T.S-., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, Vol.4, pp.1633-1649.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, Vol.32, No.3, pp.1037-1049.

Levy, E.D., Michnick, S.W., and Landry, C.R. (2012). Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philosophical Transactions of the Royal Society B*, Vol.367, pp.2594-2606.