Advances in Biological Sciences Research (ABSR), volume 6
2017 2nd International Conference on Biological Sciences and Technology (BST 2017)

Comparison of Protein Phosphorylation Site Prediction Tools

Yao SUN, Guo-qing HUANG, Yao LI, Jia-ying XUE, Qiong WU and Lei WANG[a,*]

Institute of Advanced Technology, Heilongjiang Academy of Sciences, Harbin, Heilongjiang, P.R. China 150020

[a]wleileiyu@163.com

*Corresponding author

Keywords: Phosphorylation, Phosphorylation site, Prediction tools, Protein kinases.

Abstract. To have a better understanding of phosphorylation site prediction methods, this article made a comparison of 4 online prediction tools and submitted protein sequences to these tools for analysis. The prediction results suggestted that they play a complementary role with verification experiments in protein phosphorylation study. The problems of datasets and technical factors will limit the application and these computational methods will benefit from bioinformatics improvement.

Introduction

Post-translational modification (PTM) is the chemical modification to the protein after translation, which may include modifying an existing functional group or introducing a new one to the amino acid side chain or N/C terminal. As the most common and important type of PTM, phosphorylation plays a crucial role in a wide range of cellular processes including cell metabolism, proliferation, differentiation, apoptosis, signal transduction, and gene expression [1-5]. This reaction refers to transferring a phosphate group from ATP to the specific residue on the substrate [6-7]. Protein phosphorylation and its reverse reaction dephosphorylation are respectively catalyzed by protein kinases and protein phosphatases. This reversible reaction is dynamic which means that phosphorylation status of the same protein varies in different temporal and spatial conditions.

Phosphorylation mainly occurred at serine(Ser/S), threonine(Thr/T), and tyrosine (Tyr/Y) residues in eukaryotic cells. The characteristic of the phosphorylation amino acid is that they all contain the free hydroxyl on the side chain, meanwhile have no net charge. The occurrence of phosphorylation will alter the electric property and protein structure which will greatly affect the biological functions of protein. Researches on the phosphorylation process will help gain a better understanding of many cellular activities as well as the relations between upstream/downstream proteins in the signal transduction network.

Phosphorylation Site Prediction

Different protein kinases can specifically interact with the corresponding residue on the substrate at certain conditions[8-9]. It is important to detect the target phosphorylated residues and identify the kinase of phosphorylation, and the common experimental means for phosphorylation sites identification include mass spectrometry [10] and western blot [11]. However, the number of potential phosphorylation site samples in the protein is so huge that it brings trouble for experiment design. In addition, in the

dynamic phosphorylation process, the low-abundance phosphopeptides are too labile to be captured. The experimental identification process is also time-consuming, labor-intensive, and expensive to perform and relevant technique still needs improvement. As a result, a series of computational prediction tools have been developed as a complementary method before the experiment. Making use of the datasets from phosphorylation databases and performing bioinformatics analysis, these computational prediction tools could help narrow down the pool of phosphorylation sites in the sequence of interest and provide guidance for the experiments on mutagenesis on the protein. Researchers can obtain the information of potential phosphorylation sites by submitting the protein sequence, which is easy to perform. The following parts will focus on 4 online prediction tools.

Netphos3.1

The Netphos3.1 server [12-13] (http://www.cbs.dtu.dk/s.ervices/NetPhos) predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. This is a functionally integrated version of NetPhos 2.0 and NetPhosK 1.0 which can perform generic and kinases specific predictions. It can predict phosphorylation sites of 17 kinases and the default prediction will display the results of the serine, threonine and tyrosine residues in the input sequences and generate a graphical output. Users can modify settings in the following ways: the target substrate residue can be limited to only one group of S/T/Y instead of all the three in the default mode; the output can show the predictions with the highest score for each residue compared to displaying all the results by default; in addition, a score threshold can be set to exclude the lower score predictions, and the default threshold is 0 which means that all the results will be retained; the output format can be changed to GFF, and finally the graphical output option can be canceled.

Fig. 1 was the output for P53_HUMAN (UniProt: P04637) displaying threonine only, selecting "display on the best prediction" with the threshold of 0.25. The table on the left presented the 22 predictions with the sequence name, position of residue, residue type, sequence context of 9 residues, score, kinase and answer for positive predictions. The scores > 0.5 indicated a positive prediction and the higher the score is, the more potential this residue is phosphorylated. An overview of the distribution of the predicted sites was shown in the protein sequence. Besides, the illustration on the right showed the score of each residue (green) and the threshold (pink). In this prediction 11 sites were outputted as "Yes".

GPS Web Server

The GPS Web Server [14-16] (http://gps.biocuckoo.org/online.php) has a simple operation interface in which 2 parts are well worth being pointed out: the kinase and the "Threshold". It can predict the phosphorylation sites of specific kinase, and the "Threshold" options include: "High", "Medium", and "Low" which can be set to control the FPR(false positive rate) in the predictions and there is also an "All" option which will generate the details of all the predictions.

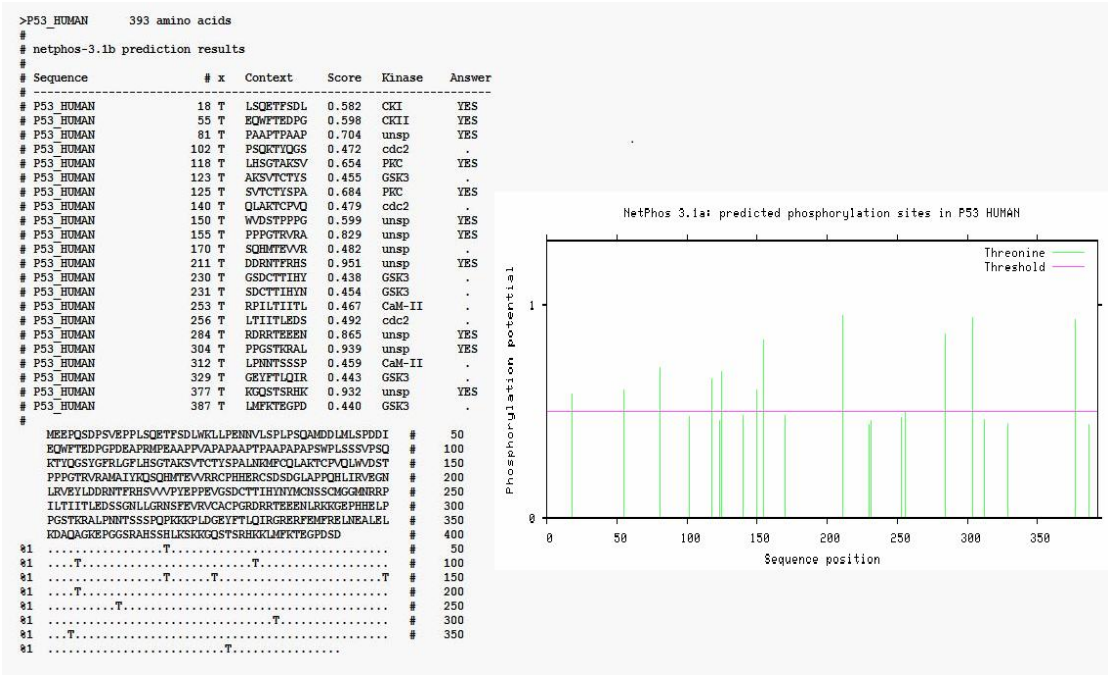

Fig.1 Illustration of Netphos3.1 predictions for53_HUMAN

As the default example, the sequence of human protein BimEL_HUMAN (UniProt: O43521-1) was submitted to predict AKT/MAPK(S/T) and Jak(Y) substrates with the "medium" option and the results were displayed in Fig. 2. The upper part in Fig 2a was the table of prediction results which included the information of protein ID, position of sites, residue, the predicted kinase type, polypeptide sequence context of 15 residues (7 residues each for upstream and downstream), GPS score and the cutoff value on the threshold. The lower part was the chart for protein disordered region predicted by IUPred [17], in which the cutoff value=0.5; that is, if score of prediction > 0.5, the residue was considered in disordered region. In Fig. 2b, not only the positions of potential phosphorylation sites were shown on the horizontal axis, but the pie chart distribution of kinases groups and S/T/Y sites as well as the bar chart distribution of the sites in the disordered region. A total of 129 items were displayed in this submission including 105 S , 21 T and 3 Y predictions, after combination of duplicate sites, these items could be classified into 15 S, 5 T and 3 Y sites. 8 S, 5 T, and 1 Y sites were predicted to be in disordered region. Another example P53_HUMAN was also analyzed under the same condition, and the results included 18 S, 9 T and 4 Y sites ,among which 16 S, 4 T and 1 Y sites were in disordered region. Users can also access the results of protein secondary structure and surface accessibility by comprehensive prediction mode, and together with the disordered region, these annotations will provide support for further analysis on the predicted sites.

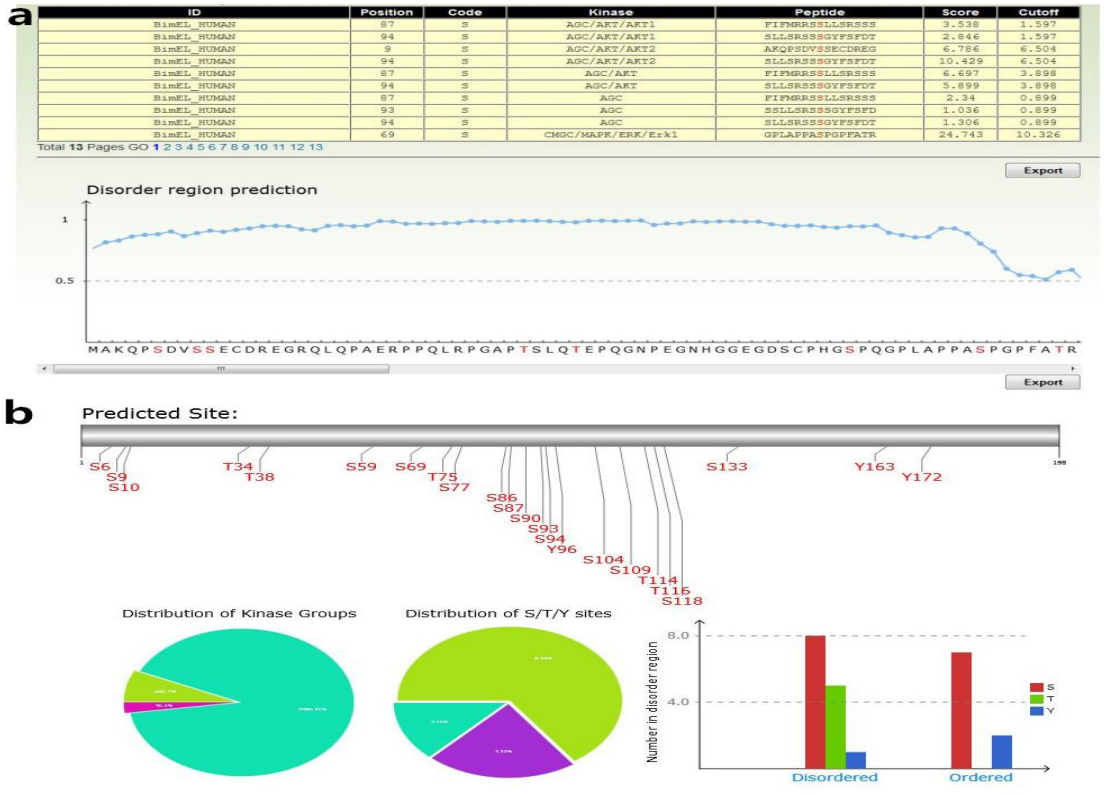

Fig. 2 Illustration of GPS Web Server predictions for BimEL_HUMAN

Fig.2a The table of the GPS 3.0 results and the visualization for protein disordered region predicted by IUPred.

Fig. 2b The distribution of the predicted sites in protein sequence and visualization by pie chart/bar chart.

PhosphoSVM

As a non-kinase-specific prediction tool, the PhosphoSVM server [18] (http://sysbio-.unl.edu/PhosphoSVM/prediction.php) only requires the protein sequence as input. While a kinase-specific prediction program requires the sequence and other relevant information including the type of kinase and the corresponding residue. It is applicable for detecting phosphorylation sites for which the kinase is unknown or the information of substrate sequence of the relevant kinase is few. With the completion of many non-model organism genome sequences, obtaining the data of more kinase species, the lack of substrate information will greatly affect the algorithms and functions of kinase-specific tools, resulting in a higher demand for non-kinase-specific tools [19]. Fig. 3 showed the results of P53_HUMAN predictions. It was well noting that the output consisted of all the S/T/Y residues (38S, 22T, and 9Y sites) in the sequence which was different from the tools mentioned above. The details of phosphorylation sites for each type of residue were shown in descending score order. Since there was no threshold option for this server, the lower score predictions could not be excluded.

Fig.3 Illustration of PhosphoSVM predictions

PhoScan

The PhosScan server [20] (http://bioinfo.au.tsinghua.edu.cn/phoscan) can predict the phosphorylation sites in two ways. One is by consensus sequences, which is applicable for 41 kinases with little phosphorylation and substrate information. Another is by log-odds ratio, which is applicable for seven kinase families. The log-odds score is defined as the log-ratio between distribution in phosphorylation sites from the positive training set and its distribution in background sites from the background set. The higher the total score on all features is, the higher likelihood this site is phosphorylated by this kinase. There are two stringency levels for log-odds ratio prediction: the high level indicates a more selective prediction; while the low level is sensitive to detecting more potential sites.

The predictions for the default example sequence in both ways were shown in Fig.4 (not fully displayed). Fig. 4a and 4b respectively showed the results of high/low stringency level. 36 sites were predicted by consensus sequences and the associated sequences were listed together with kinase type and position in the sequence. Besides, it was clear that by log-odds ratio more potential sites were detected in low stringency prediction compared to high level (17 to 5), These results demonstate that high stringency prediction could ensure the confidence of results and decrease FPR to the maximum extent.

Discussions

Computational prediction has become an advancing method for phosphorylation study and it can provide useful information before experimental verification. However, there are still some problems caused by datasets and technical issues. Protein phosphorylation researches are mainly focusing on medical and animal area, which means that compared to human and animal, the number of datasets on plant is quite limited. Some online prediction tools have used human phosphorylation databases as the training sets, for this reason, these tools are not applicable for plant phosphorylation analysis. To solve this problem, more researches on plant phosphorylation are required to enrich the plant datasets, or researchers should consider a prediction tool covering more species.

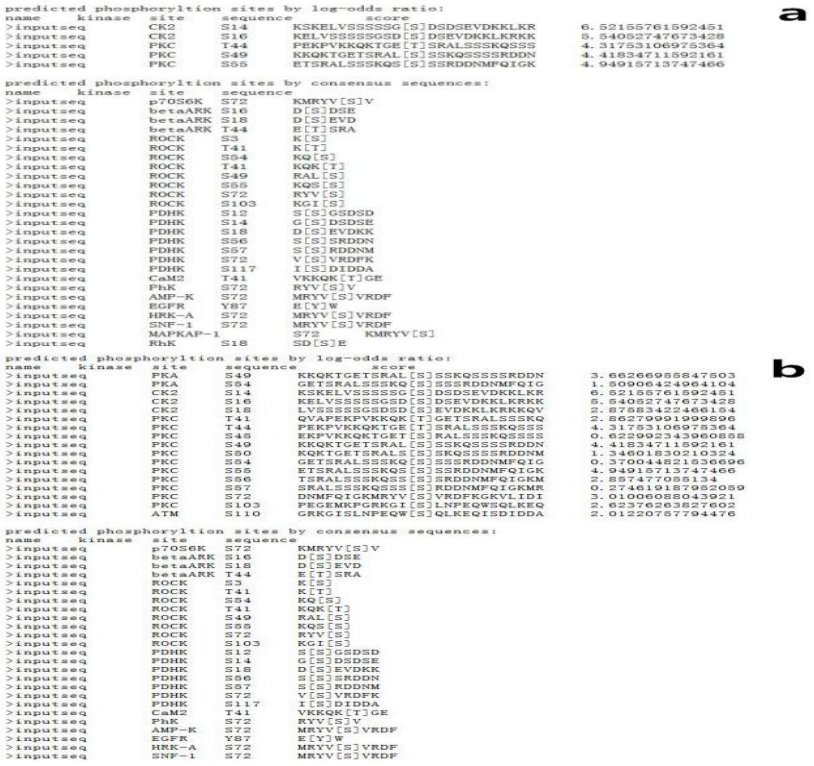

Fig.4 Illustration of PhoScan predictions

Fig.4a High stringency level. Fig.4b Low stringency level

In this article P53_HUMAN was analyzed by 3 different tools and the predictions on number and position of phosphorylation sites had a significant difference. This is because these tools are based on different algorithms and different databases. The results of P53_HUMAN predictions indicate that predicting with different tools will have a great impact further experiments. In addition, some prediction tools will list the results of all the S/T/Y sites, which is unnecessary because the number of phosphorylated sites in nature is quite limited. To exclude some of the low score predictions, the threshold and cutoff value can be introduced to the method. There is also the problem of data redundancy in some cases, that is, one potential site may be repeatedly analyzed for different kinase types, and actually some of the predictions with low confidence can be ignored. Futhermore, many unreported phosphorylated sites in current databases may be phosphorylatable, and these sites may join the negative test set, which will slightly affect the overall sensitivity and specificity. To sum up, computational predictions mainly rely on the primary structure of protein, and researchers cannot simply take it as real circumstance.

Conclusions

Previous researches concerning phosphorylation prediction mainly decribe the result of one prediction tool. In this article, we attempt to make a comparison of 4 online phosphorylation prediction tools on the operation, interfaces, functions and prediction results. It can be concluded that some prediction tools are developed from human or animal datasets, limiting the pricision in other species. Besides, the prediction differences caused by algorithms and training sets will affect the identification process. In particular, the phosphorylation process in vivo is complicated, which should not be

simply reflected by predictions from the amino acid sequence and primary structure, in short, experiments are still necessary for phosphorylation sites confirmation.

Acknowledgement

This work was supported by state key laboratory of tree genetics and breeding Northeast Forestry University, under Grant No.K2013102; and national natural science foundation under Grant No.31400576.

References

[1]T.E. Thingholm, O.N. Jensen, M.R. Larsen, Proteomics. 9(2009)1451-1468.

[2]J.J. Duan,A.F. Lozada ,C.Y. Gou , et al, Mol Cell Neurosci. 68(2015)340-349.

[3]D. Linke, T. Koudelka, A. Becker, et al, Rapid Commun Mass Sp. 29(2015):919-926.

[4]S.K. Binz , A.M. Sheehan, M.S. Wold, DNA Repair (Amst) .3(2004) 1015-1024.

[5]D.M. Clifford, S.M. Marinco, G.S. Brush, J Biol Chem, 279(2004)6163-6170.

[6]E.H. Fischer, E.G. Krebs, J Biol Chem. 216 (1955)121-132.

[7]A. Krupa, G. Preethi, N. Srinivasan, J Mol Biol. 339 (2004) 1025-1039.

[8]R.I. Brinkworth, R.A. Breinl, B. Kobe, Proc Natl Acad Sci USA .100(2003)74-79.

[9]L. Li, E.I. Shakhnovich, L.A. Mirny, Proc Natl Acad Sci USA. 100(2003)4463-4468.

[10]S.B. Breitkopf, J.M. Asara, Curr Protoc Mol Biol.2012 1-27.

[11]P. Cui, T. Chen, T. Qin, et al, Plant Cell.28 (2016)770-785.

[12]N. Blom, S. Gammeltoft, S. Brunak, J Mol Biol. 294(1999)1351-1362.

[13] N. Blom, T. Sicheritzpontén, R. Gupta, et al, Proteomics. 4(2004) 1633-1649.

[14]Y. Xue, J. Ren, X. Gao, et al, Mol Cell Proteomics. 7(2008)1598-1608.

[15]Y. Xue, F. Zhou, M. Zhu, et al, Nucleic Acids Res. 33(2005)184-187.

[16]Y. Xue, Z. Liu, J. Cao, et al, Protein Eng Des Sel. 24(2011)255-260.

[17]Z. Dosztanyi, V. Csizmok V, P. Tompa, et al, Bioinformatics. 21 (2005) 3433-3434.

[18]Y. Dou, B. Yao, C. Zhang, Amino Acids.46(2014)1459-1469

[19]B. Trost, A. Kusalik, Bioinformatics. 27(2011)2927-2935.

[20]T. Li, F. Li, X. Zhang, Proteins. 70(2008)404-414.