Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 157 (2019) 25–30

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12-13 September 2019

An Evaluation of Deep Neural Network Performance on Limited Protein Phosphorylation Site Prediction Data

Favorisen Rosyking Lumbanraja[a,*], Bharuno Mahesworo[b], Tjeng Wawan Cenggoro[b,c], Arif Budiarto[b,c], Bens Pardamean[b,d]

[a]*Department of Computer Science, Faculty of Mathematics and Natural Science, University of Lampung, Jalan Prof. Dr. Sumantri Brojonegoro No.17, 35145, Bandar Lampung, Indonesia*

[b]*Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia 11480*

[c]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

[d]*Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

One of the common and important post-translational modification (PTM) types is phosphorylation. Protein phosphorylation is used to regulate various enzyme and receptor activations which include signal pathways. There have been many significant studies conducted to predict phosphorylation sites using various machine learning methods. Recently, several researchers claimed deep learning based methods as the best methods for phosphorylation sited prediction. However, the performance of these methods were backed up with the massive training data used in the researches. In this paper, we study the performance of simple deep neural network on the limited data generally used prior to deep learning employment. The result shows that a deep neural network can still achieve comparable performance in the limited data settings.

Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2019.

Keywords: Phosphorylation Site Prediction; Protein Phosphorylation; Deep Learning; Deep Neural Network

1. Introduction

One of the most important post-translational modifications [PTMs] is phosphorylation. With protein kinase, protein phosphorylation occurs when a phosphate group is added to an amino acid. These amino acids are Serine (S), Threonine (T), dan Tyrosine (Y)[1]. It is also the most used post-translational modification in eukaryotes[2 3] and play crucial

* Corresponding author. Tel.: +62-812-8608-197.
E-mail address: favorisen.lumbanraja@fmipa.unila.ac.id

1877-0509 © 2019 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2019.
10.1016/j.procs.2019.08.137

roles in many cellular behaviour, such as metabolism[4], DNA repair[5], environmental stress response[6], regulation of transcription[7], and other important processes[8]. Therefore abnormality in protein can affect these cellular processes which may lead to many kinds of diseases. Because of that reason, it is important to identify and learn more about phosphorylation in the cell.

In general, there are two approaches to predict phosphorylation sites. The first approach is the kinase-specific phosphorylation prediction site. This approach requires information about the protein sequences, which are phosphorylated kinase enzymes. However, the main constraint of this approach is that kinase enzyme information for the public is limited[9]. The second approach is non-kinase-specific phosphorylation prediction site. This approach only requires the information of the phosphorylated protein sequences to predict the phosphorylated site. The differences and comparisons between these two approaches are explained by Xue et al. in their paper[10].

The typical studies of phosphorylation site prediction heavily employ machine learning algorithm as a site predictor. Following the recent trend in machine learning, recent studies of phosphorylation site prediction use deep learning with massive training dataset[11,12]. However, prior to 2017, the training dataset used is relatively small and limited to only 9 amino acids per sequence. Therefore, we cannot fairly compare the performance of deep learning based method with other well-established methods. In this study, we explore the performance of a simple deep neural network in the limited data settings typically used before 2017. We argue that, with the recent invention of various techniques in deep learning, a simple deep neural network is capable to achieve powerful performance even with limited training data.

2. Related Works

Since the advent of deep learning, the trend of machine learning research is shifted to the utilization of dataset with massive size. The emergence of this trend is due to the exceptional performance of deep learning given a massive training dataset. This trend is also apparent in phosphorylation site prediction. For instance, Musite Deep[11] and Deepphos[12] used dataset with 913,623 and 335,622 sites respectively.

The studies prior to Deep Musite typically used P.ELM[13] and PPA[14], which has only 4,750 and 852 sites. These studies generally employed popular machine learning models at that time, such as SVM[15,16,17,18] and Random Forest[15,19,20]. These models have limited performance on large data, thus they reduce the data further by cropping the protein sequences to 9 amino acids instead of using full sequence.

3. Materials and Methods

3.1. Materials and Datasets

The datasets used in this study are composed of polypeptide sequences, where each sequence consisting of 9 amino acid. We define this fixed length with 9 amino acid as window 9 sequence. The fifth amino acid or the amino acid in the middle of the sequence is the amino acid with the possible location for phosphorylation, Serine (S), Threonine (T), or Tyrosine (Y). Each sequence is labelled as a positive or negative sequence, where positive means that a phosphorylation event occurs on that location.

The window 9 sequence is generated from P.ELM database version 9[13] and PPA database[14]. The sequences are grouped according to their database source and phosphorylatable residues (Serine, Threonine, or Tyrosine). To reduce redundancy, sequences with the similarity of 0% to 20%, gap opening set the value to 10, and also the value of gap extension to 0.5, were removed using skipredundant[21]. We used the exact same datasets that were used by Lumbanraja et al[15] in their study. Table 1 below shows the size of each dataset.

3.1.1. Neural Network Architecture

The neural network model used in this study consists of 4 fully-connected layers. Prior to each fully-connected layer, we use Batch Normalization[22] layer to stabilize the training of the model. Every Dense layer has 32 neurons and exponential linear unit (ELU)[23] as the activation functions. We use an embedding model as a feature representation for each amino acid. The embedding model is inspired by word2vec model[24]. We modify word2vec model to encode each amino acid as a vector instead of words. The architecture of our neural network is illustrated in Figure 1.

Table 1: Datasets size.

Datasets	Serine		Threoninie		Tyrosine	
	Positive	Negative	Positive	Negative	Positive	Negative
P.ELM	1554	1543	707	453	267	226
PPA	307	307	68	68	51	51

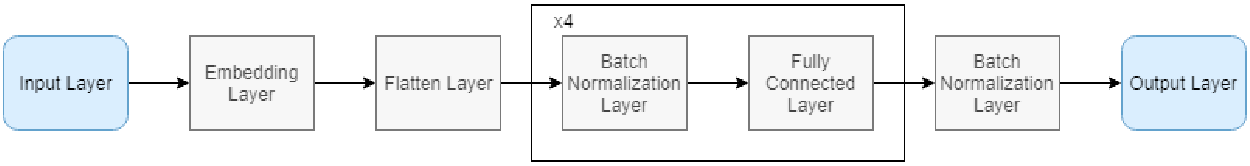

Fig. 1: Neural Network Diagram

To optimize the model, we use Adam optimizer[25] with standard configuration to optimize the model. The model is trained for 100 epochs with a scheduled learning rate decrease. The learning rate starts at 0.001, then it is reduced to 0.0005, 0.0002, and 0.0001 subsequently at 10th, 40th, and 70th epoch. To decrease overfitting, we utilize Dropout[26] with a drop rate of 0.1 and l2 regularization with a rate of 0.0001.

3.1.2. Evaluation

To evaluate our model, 10-folds cross-validation method is used to score our phosphorylation site prediction algorithm. Our datasets are split into 10 folds proportionately. Afterwards, we train and evaluate the model ten times, with each fold is used as validation split in turns.

To measure the performance of our method, we use the same metrics as in the study done by Lumbanraja et al[15]: Accuracy, Sensitivity, Specificity, F1 score, Area Under Curve (AUC)[27], and Matthews Correlation Coefficient (MCC). The metrics are formulated as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$F1score = \frac{TP}{TP + FP + FN} \tag{4}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

After the neural network is trained and validated using the metrics above, we took the average score of the 10 validation scores to compare with the previous methods.

4. Result and Discussion

The performance of our proposed method is shown in Table 2 below. It can be seen that the performance is generally proportional to the size of the dataset. We can see that the performance of the neural network on P.ELM Serine dataset which is the biggest dataset, gave the best result. On the other hand, our deep neural network gave the lowest result

Table 2: Proposed Method's Performance.

Metrices	P.ELM			PPA		
	Serine	Threonine	Tyrosine	Serine	Threonine	Tyrosine
Dataset Size	3097	1160	493	614	136	102
Accuracy	0.9146	0.8733	0.7564	0.8109	0.8242	0.6409
AUC	0.9185	0.8708	0.7602	0.8104	0.8288	0.6565
Sensitivity	0.9305	0.8768	0.7272	0.8247	0.8034	0.6536
Specificity	0.9065	0.8648	0.7933	0.7962	0.8542	0.6595
F1-Score	0.9197	0.8939	0.7609	0.8111	0.8076	0.6339
MCC	0.8385	0.7362	0.5186	0.6211	0.6597	0.3120

on the PPA Tyrosine dataset, which is the smallest dataset. Therefore, we can conclude that, despite the limited data setting, the dataset size still plays an important role in deep neural network performance.

Currently, the best method for phosphorylation site prediction in limited data is feature extraction and feature selection developed by Lumbanraja et al [15]. Therefore, we compare our deep neural network to the Lumbanraja et al. method along with other well-established phosphorylation site prediction methods: Netphos K [28], GPS 2.1 [10], Swaminathan et al. method [17], Netphos [29], PPRED [30], PHOSFER [20], Musite [18], Phospho SVM [16], and RF-Phos [19]. We do not compare our method with another deep learning based method such as Musite Deep [11] and Deepphos [12], because these methods are tailored for full sequence setting instead of window 9 sequence setting. The comparison is shown in Table 3 and 4 for both datasets P.ELM and PPA. The best performance is marked with ***bold-italic*** font, while for the second best is marked with **bold** font. We can see that our simple deep neural network can deliver a comparable performance among the other methods. It is even the second best method overall behind Lumbanraja et al. method [15]. This suggests that a simple deep neural network can still deliver a powerful performance despite using limited training data.

Table 3: Result Comparison for P.ELM Datasets.

Method	Serine				Threonine				Tyrosine			
	AUC	Sens	Spec	MCC	AUC	Sens	Spec	MCC	AUC	Sens	Spec	MCC
Netphos K	0.63	0.51	0.68	0.08	0.6	0.62	0.57	0.07	0.6	0.39	0.74	0.08
GPS 2.1	0.73	0.33	0.93	0.2	0.7	0.38	0.93	0.2	0.61	0.34	0.79	0.08
Swaminathan et al.	0.7	0.31	0.89	0.13	0.72	0.28	0.92	0.14	0.62	0.60	0.57	0.09
Netphos	0.7	0.34	0.87	0.12	0.66	0.34	0.84	0.09	0.65	0.35	0.84	0.13
PPRED	0.75	0.32	0.92	0.17	0.73	0.30	0.91	0.13	0.7	0.43	0.83	0.17
Musite	0.81	0.41	**0.94**	0.25	0.78	0.34	***0.95***	0.22	0.72	0.384	**0.87**	0.18
Phospho SVM	0.84	0.44	**0.94**	0.3	0.82	0.38	***0.95***	0.25	0.74	0.42	**0.87**	0.21
Rf-Phos	0.88	0.84	0.85	0.65	**0.9**	0.83	**0.94**	0.7	***0.91***	**0.83**	***0.88***	***0.7***
Lumbanraja et al.	***0.96***	***0.97***	***0.96***	***0.93***	***0.92***	***0.93***	0.92	***0.84***	**0.8**	***0.84***	0.76	**0.6**
Our Method	**0.92**	**0.93**	0.91	**0.84**	0.87	**0.88**	0.86	**0.74**	0.76	0.73	0.79	0.52

5. Conclusions

In this paper, we show that a simple deep neural network can achieve comparable performance to other state-of-the-art models for phosphorylation site prediction with limited data. This study suggests that the remarkable performance of deep learning in phosphorylation site prediction is not only due to the massive dataset used for training. Therefore, the use of more complex deep learning method is suggested for future study in phosphorylation site prediction.

Table 4: Result Comparison for PPA Datasets.

Method	Serine			Threonine			Tyrosine		
	Sens	Spec	MCC	Sens	Spec	MCC	Sens	Spec	MCC
Netphos K	0.80	0.39	0.10	0.69	0.51	0.06	0.25	**0.83**	0.04
GPS 2.1	***0.95***	0.29	0.14	***0.96***	0.21	0.07	***0.98***	0.21	0.09
Netphos	0.77	0.54	0.16	0.54	0.77	0.12	0.65	0.67	0.13
PHOSFER	0.75	0.66	0.22	0.78	0.65	0.14	0.63	0.59	0.08
Musite	0.56	***0.87***	0.31	0.49	***0.94***	0.26	0.47	***0.89***	0.20
Phospho SVM	0.64	0.81	0.29	0.71	0.82	0.19	**0.82**	0.64	0.18
Rf-Phos	0.72	0.70	0.41	0.79	0.70	0.50	0.61	0.62	**0.29**
Lumbanraja et al.	**0.89**	**0.86**	***0.76***	**0.88**	***0.94***	***0.82***	0.53	0.63	0.16
Our Method	0.82	0.80	**0.62**	0.80	**0.85**	**0.66**	0.65	0.66	***0.31***

Acknowledgements

The experiment in this research used NVIDIA Tesla P100 provided by NVIDIA - BINUS AI R&D Center. This research is a collaboration between Department of Computer Science, University of Lampung and Bina Nusantara University.

References

1. Hunter, T.. Signaling-2000 and Beyond. *Cell* 2000;**100**:113–127.
2. Khoury, G., Baliban, R., Floudas, C.. Proteome-wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot database. *Scientific Report* 2011;**1**(9).
3. Pinna, L., Ruzzene, M.. How Do Protein Kinases Recognize Their Substrates? *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1996;**1314**(3).
4. Bu YH He YL, Z.H.L.W.P.D.T.A.T.L.X.H.H.Q.L.X.L.E.. Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. *Journal of Endocrinology* 2010;**206**:271–277.
5. C. David Wood Tina M. Thornton, G.S.R.A.D., Rincon, M.. Nuclear localization of p38 mapk in response to dna damage. *International Journal of Biological Sciences* 2009;**5**(5):428–437.
6. Ye-yu Wang, S.m.C..H.L.. Hydrogen peroxide stress stimulates phosphorylation of foxo1 in rat aortic endothelial cells. *Acta Pharmacologica sinica* 2010;**31**:160–164.
7. Uddin, S., Lekmine, F., Sassano, A., Rui, H., Fish, E.N., Platanias, L.C.. Role of stat5 in type i interferon-signaling and transcriptional regulation. *Biochemical and Biophysical Research Communications* 2003;**308**(2):325 – 330. URL http://www.sciencedirect.com/science/article/pii/S0006291X03013822.
8. Trost, B., Kusalik, A.. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* 2011;**27**(21):2927–2935. http://oup.prod.sis.lan/bioinformatics/article-pdf/27/21/2927/583531/btr525.pdf; URL https://doi.org/10.1093/bioinformatics/btr525.
9. Newman, R.H., Hu, J., Rho, H.S., Xie, Z., Woodard, C., Neiswinger, J., et al. Construction of Human Activity-based Phosphorylation Networks. *Molecular Systems Biology* 2013;**9**(1).
10. Xue, Y., Liu, Z., Cao, J., Ma, Q., Gao, X., Wang, Q., et al. GPS 2.1: Enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Engineering, Design and Selection* 2011;**24**(3):255–260.
11. Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., et al. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 2017;**33**(24):3909–3916.
12. Luo, F., Wang, M., Liu, Y., Zhao, X.M., Li, A.. Deepphos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics* 2019;.
13. Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., et al. Phospho.ELM: A database of phosphorylation sites-update 2011. *Nucleic Acids Research* 2011;**39**(SUPPL. 1):261–267.
14. Heazlewood, J.I., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., et al. PhosPhAt : A database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Research* 2008;**36**(SUPPL. 1):1015–1021.
15. Lumbanraja, F.R., Nguyen, N.G., Phan, D., Faisal, M.R., Abapihi, B., Purnama, B., et al. Improved Protein Phosphorylation Site Prediction by a New Combination of Feature Set and Feature Selection. *Journal of Biomedical Science and Engineering* 2018;**11**(06):144–157. URL http://www.scirp.org/journal/doi.aspx?DOI=10.4236/jbise.2018.116013.
16. Dou, Y., Yao, B., Zhang, C.. PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* 2014;**46**(6):1459–1469.

17. Swaminathan, K., Adamczak, R., Porollo, A., Meller, J.. Enhanced prediction of conformational flexibility and phosphorylation in proteins. In: *Advances in Computational Biology*. Springer; 2010, p. 307–319.
18. Gao, J., Thelen, J.J., Dunker, A.K., Xu, D.. Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. *Molecular & Cellular Proteomics* 2010;**9**(12):2586–2600. URL http://www.mcponline.org/lookup/doi/10.1074/mcp.M110.001388.
19. Ismail, H.D., Jones, A., Kim, J.H., Newman, R.H., Kc, D.B.. RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest. *BioMed Research International* 2016;**2016**.
20. Trost, B., Kusalik, A.. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* 2013;**29**(6):686–694.
21. Sikic, K., Carugo1, O.. Protein sequence redundancy reduction: comparison of various methods. *Bioinformation* 2010;**5**(6):234–239.
22. Ioffe, S., Szegedy, C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015, .
23. Djork-Arne Clevert, , Unterthiner, T., Hochreiter, S.. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUS). *ICLR 2016* 2016;1511.07289v5.
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. 2013, p. 3111–3119.
25. Kingma, D.P., Ba, J.L.. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*. 2015, .
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Ruslan Salakhutdinov, . Dropout: A Simple Way to Prevent Neural Networks from Overfitting 2004;**1**(60):11. 1102.4807.
27. Bradley, A.P.. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997;**30**(7):1145 – 1159. URL http://www.sciencedirect.com/science/article/pii/S0031320396001422.
28. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., Brunak, S.. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 2004;**4**(6):1633–1649.
29. Blom, N., Gammeltoft, S., Brunak, S.. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology* 1999;**294**(5):1351–1362.
30. Biswas, A.K., Noman, N., Sikder, A.R.. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics* 2010;**11**.