DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases

Iman Deznabi [1 2] **Busra Arabaci** [1] **Mehmet Koyut** [3 4] **Oznur Tastan** [5]

Abstract

Kinases catalyze the phosphorylation of other proteins in a target specific manner. The dysregulation of phosphorylation is associated with many diseases, including cancer. Determining which kinase is responsible for phosphorylating a site remains an experimental challenge. Existing computational methods require several examples of known targets of a kinase to make accurate kinase specific predictions, yet for a large body of kinases, only a few or no target sites are reported. We recently presented DeepKinZero (), the first zero-shot learning approach to predict the kinase acting on a phosphosite for kinases with no known phosphosite information through a zero-shot learning model. We showed that DeepKinZero achieves significant improvement in accuracy for kinases with no known phosphosites in comparison to the baseline model and other methods available. By expanding our knowledge on understudied kinases, DeepKinZero is available at https://github.com/Tastanlab/DeepKinZero.

1. Introduction

Phosphorylation is the key mechanism for regulating protein function in signal transduction (Hunter, 1995). The amino acid residue that receives the phosphoryl group is usually called the phosphorylation site, or briefly a *phosphosite*. Kinases are the enzymes that catalyze the phosphorylation event. Due to their central role in a broad range of cellular activities, aberrant kinase function is implicated in many diseases, and they are a major class of drug targets (Ferguson & Gray, 2018).

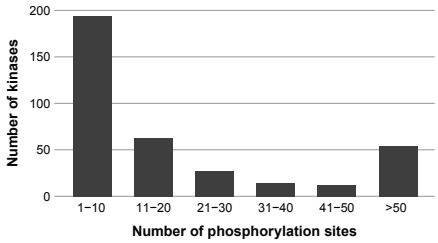

Figure 1. The histogram of the number of experimentally validated target phosphosites for human kinases in PhosphositePlus.

Although advances in phosphoproteomics enable identification of phosphosites at the proteome level, most of the phosphoproteome is still in the dark: more than 95% of reported human phosphosites have no known kinase or associated biological function (Needham et al., 2019). Identifying the kinase that is responsible for catalyzing a phosphosite is a key question. With 518 identified kinases in the human genome and the transient nature of kinase-substrate interactions, it is experimentally challenging to determine the cognate kinase of a phosphosite.

Most of the earlier computational models employ supervised machine learning models; thus, the application of such tools is limited to kinases for which a substantial number of target phosphosites are available for training. However, for a large body of kinases, no or only a few target sites are reported (Figure 1). For example, MusiteDeep (Wang et al., 2017), uses deep learning to predict binding sites for kinases, and it exclusively focuses on kinase families with at least 100 experimentally verified phosphosites. Recently, we presented DeepKinZero (Deznabi et al., 2020), the first zero-shot learning approach to predict the kinase acting on a phosphosite for kinases with no known phosphosite information.

2. Methods

Zero-shot learning aims at solving classification problems wherein the available training data does not contain examples of the desired classes (Akata et al., 2016). The key to making predictions for classes with no training data (referred to as *unseen* or *zero-shot* classes) is to have side

[1]Computer Engineering Department, Bilkent University, Ankara, Turkey [2]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA 01003, USA [3]Dept of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106, USA [4]Center for Proteomics & Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA [5]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, 34956, Turkey. Correspondence to: Oznur Tastan <otastan@sabanciuniv.edu>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

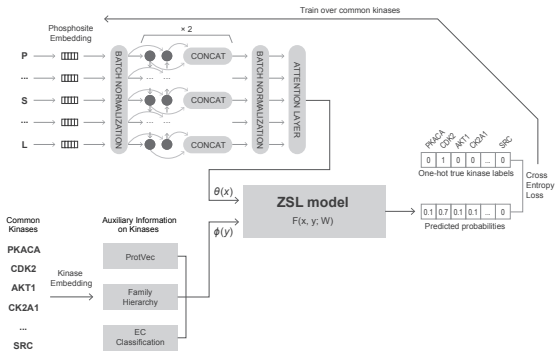

Figure 2. DeepKinZero model architecture.

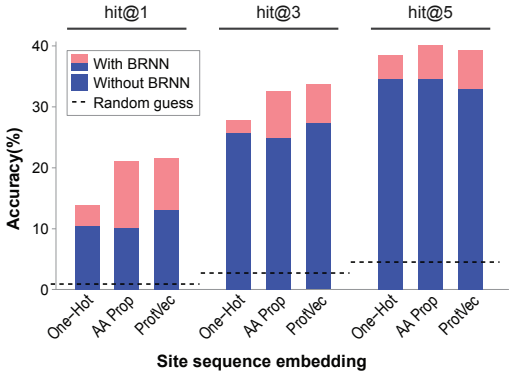

Figure 3. Performance comparison of the models with the site sequence embeddings and with and without using a BRNN.

information which can be used to relate the classes. Based on these relations among classes, it becomes possible to transfer the knowledge obtained from classes that have positive training samples (referred to as *seen* class) to the previously unseen classes (Akata et al., 2016). Thus, even if we do not observe any phosphosites that are associated with a rare kinase (unseen class) in training, the zero-shot learning framework enables us to recognize a target site of this kinase by transferring knowledge from common kinases to the rare kinases. This can be achieved by establishing a relationship between the kinases using relevant auxiliary information, such as functional, sequence, and structural characteristics of kinases.

The problem is formalized as a multi-class classification problem, where each input phosphosite sequence is associated with a kinase. The overall architecture is provided in (Figure 2). DeepKinZero takes the sequence of 15 residues centered on the phosphosite as input, x. For each phosphosite $x \in \mathcal{X}$, we compute phosphosite embedding vector, $\theta(x) \in \mathbb{R}^d$, that represents the phosphosite sequence in a d-dimensional space. To learn this embedding, we use two layers of Bidirectional Recurrent Neural Networks(BRNN) followed by a dot attention layer over phosphosite embeddings.

For each kinase $y \in Y$, a "kinase embedding" vector $\phi(y) \in \mathbb{R}^m$ is computed based on information available on kinases. We use four different data sources to represent kinases based on their sequence, kinase family taxonomy, enzyme classification and the pathways they participate. We expect "similar" kinases to be close according to the Euclidean metric in the embedded space.

For the zero-shot learning, as in (Sumbul et al., 2018), we define a compatibility function F as $F(x, y) = [\theta(x)^\top \quad 1]^\top W [\phi(y)^\top \quad 1]$, which takes phosphosite - kinase pair, (x, y), as input and gives a scalar value proportional to the probability of associating the site x with kinase y. Here, W denotes the $(d + 1) \times (m + 1)$ compatibility matrix. We learn W by minimizing the cross-entropy loss over the training data.

3. Results

We train and evaluate our models on the experimentally validated kinase-phosphosite associations obtained from PhosphoSitePlus (Hornbeck et al., 2014). Following the evaluation protocol suggested in (Xian et al., 2017), we split the data into training, validation and test data based on the number of sites that are associated with each kinase. We use hit@k accuracy on the test data to evaluate performance.

Figure 3 summarizes the results of using different phosphosite sequence embeddings. In hit@1 and hit@3 metrics, site sequence embeddings obtained using a BRNN coupled with ProtVec vectors performs the best. In hit@5, input representations do not yield to very different results, and indeed the amino acid properties with BRNN have the highest hit@5 accuracy. Additionally, we observe that the use of BRNN model significantly improves the performance. Where the model using BRNN and ProtVec site embeddings have the highest performance with 21.52% test accuracy. Note that these numbers are highly impressive because it would not be possible to train predictive models for these kinases due to the inadequacy of training samples, and random guess will achieve only 0.89% accuracy since there are 112 test classes. We also evaluated different combinations of kinase embedding (Deznabi et al., 2020). Among the four possible kinase embeddings, the kinase hierarchy of kinases contributes the most to the accuracy of the model, achieving 17.72% accuracy when used as the sole auxiliary information on kinases. Inspection of model weights reveal that the model correctly learns to assign more weight to the center(Deznabi et al., 2020). The model also successfully learns the amino acid preferences of the kinase families.

4. Conclusion

DeepKinZero is a novel method for kinase-specific phosphorylation site predictions. DeepKinZero, unlike conventional supervised methods can offer predictions for kinases that do not have any known phosphosites, which will be helpful in illuminating the dark phosphoproteome.

References

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2016.

Deznabi, I., Arabaci, B., Koyutürk, M., and Tastan, O. DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases. *Bioinformatics*, 02 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa013. URL https://doi.org/10.1093/bioinformatics/btaa013. btaa013.

Ferguson, F. M. and Gray, N. S. Kinase inhibitors: the road ahead. *Nature Reviews Drug Discovery*, 17(5):353, 2018.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2014.

Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*, 80 (2):225–236, 1995.

Needham, E. J., Parker, B. L., Burykin, T., James, D. E., and Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.*, 12(565):eaau8645, 2019.

Sumbul, G., Cinbis, R. G., and Aksoy, S. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2018.

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 33(24):3909–3916, 2017.

Xian, Y., Schiele, B., and Akata, Z. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.