PhosphoGAN: Enhancing the prediction process of general and kinase-specific phosphorylation sites


**Document Version:**
Peer reviewed version

**Queen's University Belfast - Research Portal:**
[Link to publication record in Queen's University Belfast Research Portal](#)

**Publisher rights**
© 2018 The Authors.
This is an open access article published under a Creative Commons Attribution-NoDerivs License (https://creativecommons.org/licenses/by-nd/4.0/), which permits reproduction and redistribute in any medium, provided the author and source are cited and any subsequent modifications are not distributed.

**General rights**
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
1. Abstract

Phosphorylation site prediction has now become a cornerstone within protein function studies and experimental design. Finding the exact location of these binding sites (aka motifs) is essential in a variety of domains such as drug design and development. To address this issue, several computational techniques have been developed in recent years that attempt to characterise and predict these sites via various forms of feature extraction from raw protein sequence data. When handling raw protein sequences, two central questions typically arise:

1. Firstly, can this particular protein sequence be phosphorylated?
2. Secondly, if this sequence can indeed be phosphorylated what specific kinase is the cause of this phosphorylation?

We present PhosphoGAN, a semi-supervised generative adversarial approach to training a deep neural network that is capable of outperforming current state of the art models such as MusiteDeep in predicting general and kinase-specific phosphorylation sites. PhosphoGAN produces a classifier via a semi-supervised approach using a generative adversarial network (GAN) to aid the training process as opposed to using such methods as transfer learning.

2. Methodology

The discriminator [3] used in PhosphoGAN begins with a set of convolutional layers (i.e. CNN). The output of the CNN is the copied and one version is transposed. Both copies are then passed through a dedicated attention mechanism. Accompanying each standalone attention mechanism, we have also introduced a pair of individual bidirectional long short-term memory (BiLSTM) models with their own independent attention mechanisms.

The two-dimensional BiLSTM attention mechanism [1] (Att-BiLSTM) will analyse the feature maps produced by the CNN in both a sequence and feature map dimension. The outputs of the both Att-BiLSTM models and the stand-alone attention mechanisms are then concatenated together to form a set of features produced by the discriminator. These features are then passed through a series of fully connected layers to attain a final classification.

3. Summary of Key Findings

The phosphorylation data used in the experiment is for Homo sapiens and was gathered from UniProt/Swiss-Prot. It consisted of the phosphorylation sites on serine (S), threonine (T) and tyrosine (Y), which provided a source for the positive data for the experiment, while the negative data was taking the same amino acid excluding annotated phosphorylation sites from the proteins.

To evaluate the performance of both models, a five cross-fold validation was used, and the area under the receiver characteristic curve, average precision and F1 scores were then calculated for each fold. Whereby PhosphoGAN outperformed MusiteDeep, and MusiteDeep-GAN in both general phosphorylation site and kinase-specific prediction.

4. Conclusion

By applying a new semi-supervised training approach along with a new model architecture for the classifier, we obtain results that outperform the current state of the art MusiteDeep model. These results demonstrate how deep learning can be applied with significant effect to a problem where the training data is insufficient and unbalanced.