Musite: Global Prediction of General and Kinase-Specific Phosphorylation Sites

Jianjiong Gao[1,2,*], Jay J. Thelen[2,3], A. Keith Dunker[4], and Dong Xu[1,2]

[1]Department of Computer Science, [2]C.S. Bond Life Sciences Center, [3]Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, [4]Center for Computational Biology and Bioinformatics, Indiana University Schools of Medicine and Informatics, Indianapolis, Indiana 46202, USA

* Email: jgao@mail.mizzou.edu

Project URL: http://musite.sourceforge.net/

Source code: http://musite.svn.sourceforge.net/viewvc/musite/musite/

Reversible protein phosphorylation is one of the most pervasive posttranslational modifications, regulating diverse cellular processes in various organisms. Since mass spectrometry-based experimental approaches for identifying phosphorylation events are costly, time consuming, and are biased towards abundant proteins and proteotypic peptides, *in silico* prediction of phosphorylation sites is an attractive alternative for whole proteome annotation. Due to various limitations, current phosphorylation-site prediction tools were not well-designed for comprehensive assessment of proteomes. Here, we present a novel software tool, Musite, specifically designed for large-scale prediction of both general and kinase-specific phosphorylation sites. We collected high confidence phosphoproteomics data from multiple organisms and used these to train prediction models by a comprehensive machine learning approach. Application of Musite on proteomes of *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana* yielded tens of thousands of phosphorylation-site predictions at a high stringency level. Cross-validation tests show that Musite significantly outperforms existing tools for predicting general phosphorylation sites and is at least comparable to those for predicting kinase-specific phosphorylation sites. Furthermore, Musite provides several other unique functionalities such as customized model training and continuous stringency selection by users. Musite provides a useful bioinformatics tool to biologists for predicting phosphorylation sites *en masse* and training prediction models from custom phosphorylation data. In addition, with its easily-extensible open-source application programming interface (API), Musite is aimed at being an open platform for community-based development of machine-learning based phosphorylation-site prediction applications. Musite is available at http://musite.sourceforge.net/.

Acknowledgement

This work was supported, in part, by the funding from the National Science Foundation-Plant Genome Research Program [grant number DBI-0604439 awarded to JJT] and the National Institute of Health [grant number R21/R33 GM078601 awarded to DX]. The authors thank Dr. Thorsten Joachims and Dr. Weizhong Li for permissions of integrating SVM^{light} and CD-HIT in Musite.

Reference

Gao, J., Thelen, J.J., Dunker, A.K., and Xu, D. Musite: a Tool for Global Prediction of General and Kinase-Specific Phosphorylation Sites. *Molecular & Cellular Proteomics*. 2010. Submitted.