Bioinformatics Advance Access published February 5, 2007

Sequence analysis

NetPhosYeast: Prediction of protein phosphorylation sites in yeast

Christian R. Ingrell[1], Martin L. Miller[2] Ole N. Jensen[1], and Nikolaj Blom[2*]

[1]University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark, [2]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Anker Engelunds Vej 1, DK-2800 Kgs. Lyngby, Denmark

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: We here present a neural network based method for the prediction of protein phosphorylation sites in yeast – an important model organism for basic research. Existing protein phosphorylation site predictors are primarily based on mammalian data and show reduced sensitivity on yeast phosphorylation sites compared to those in humans, suggesting the need for a yeast-specific phosphorylation site predictor. NetPhosYeast achieves a correlation coefficient close to 0.75 with a sensitivity of 0.84 and specificity of 0.90 and outperforms existing predictors in the identification of phosphorylation sites in yeast.

Availability: The NetPhosYeast prediction service is available as a public web server at http://www.cbs.dtu.dk/services/NetPhosYeast/

***Contact**: nikob@cbs.dtu.dk

1 INTRODUCTION

Protein phosphorylation is a post-translational modification catalyzed by protein kinases. Reversible protein phosphorylation is a universal regulatory mechanism of multiple cellular functions in the Eukaryote, Prokaryote and Archaea kingdom. As the number of sequenced genomes rapidly increases, the functional annotation of the gene products is lacking behind. Since computational analysis of a protein sequence is often the first step toward understanding its function, it is important to develop and improve such computational methods. Several algorithms for predicting protein phosphorylation from the amino acid sequence are available, including Scansite 2.0 (Obenauer et al. 2003), Prosite (Sigrist et al. 2002), Netphos (Blom et al. 1999), Netphosk (Blom et al. 2004), GPS (Xue et al. 2005), Disphos (Iakoucheva et al. 2004), kinasePhos (Huang et al. 2005), PPSP (Xue et al. 2006). None of these methods are directed to predict yeast phosphorylation sites, and rely primarily on annotated phosphorylation sites identified with classical low through-put biochemical experiments extracted from databases such as Phospho.ELM (Diella et al. 2004) and PhosphoBase (Blom et al. 1998). Advances in biological mass spectrometry has enabled identification of hundreds of protein phosphorylation sites in a single experiment (Jensen 2006).

Recently, two large-scale phosphoproteomic studies mapped more than 900 phosphorylation sites in yeast (Ficarro et al. 2002; Gruhler et al. 2005), providing the foundation to develop a predictor for yeast protein phosphorylation. Although many protein kinases in yeast have homologues in humans, and vice versa, many kinases are not shared between these species. An evolutionary study of protein kinases showed that 32 kinases are unique in yeast covering unicellular functions such as osmotic and stress response, cell wall signalling, cell-cycle regulation , and small molecule transport (Ball et al. 2000). Similarly, humans have protein kinases governing development, differentiation and intercellular communication (Manning et al. 2002) that are not found in yeast. We here present the first yeast-specific phosphorylation predictor with high sensitivity and specificity. We also demonstrate that existing predictors, which are based primarily on mammalian phosphorylation sites, exhibit lower performance on known phosphorylation sites in yeast proteins. The yeast-specific protein phosphorylation site predictor, NetPhosYeast, will facilitate more confident computational annotation of yeast proteins.

2 METHODS

We generated a positive data set consisting of yeast serine and threonine phosphorylation sites experimentally identified by mass spectrometry driven phosphoproteomics from (Gruhler et al. 2005) and (Ficarro et al. 2002). We also included annotated phosphorylation sites from the Swiss-Prot database constrained not to include the modifiers “potential”, “probable”, “by similarity” or “autocatalysis” in the description field. After merging of the three data sets, redundant 7-mer phosphopeptides were removed. This yielded a total of 953 phosphoserine sites and 192 phosphothreonine sites from 675 yeast proteins. The negative data was compiled by randomly collecting non-phosphorylated serines and threonines in yeast phosphoproteins. For comparison, 1696 annotated human serine and threonine phosphorylation sites were extracted from the Swiss-Prot database not including sites with the modifier “potential”, “probable”, “by similarity” or “autocatalysis” in the description field.

Prior to training the artificial neural networks (ANNs) the negative and positive dataset were pooled. N-fold cross-validation (Nielsen et al. 2003) is typically used to estimate the accuracy of a machine learning scheme. In n-fold cross-validation the pooled dataset is partitioned into a number of subsets, including one test set and a number of training sets. Using this strategy the ANN training is performed by shifting the test set stepwise so all data is used for training and test when completed. For each test set a number of ANN parameters (window size and number of hidden neurons) are optimized according to the Matthews correlation coefficient (MCC) and an optimal parameter space is chosen. We devised a new evaluation scheme for the n-Fold cross-validation procedure. In our scheme the cross-validation procedure is extended from traditionally using a test and training set to also include an evaluation set. In this approach the pooled data set is divided into 5 sub-

[*]To whom correspondence should be addressed.

sets by random partitioning. For each subset 4-fold cross-validation is performed, but instead of using the test-set performance we calculate the performance based on the respective evaluation set. Thus, we obtain a fair and independent performance estimate of our ANN. We suggest that this training method should be termed cross-evaluation.

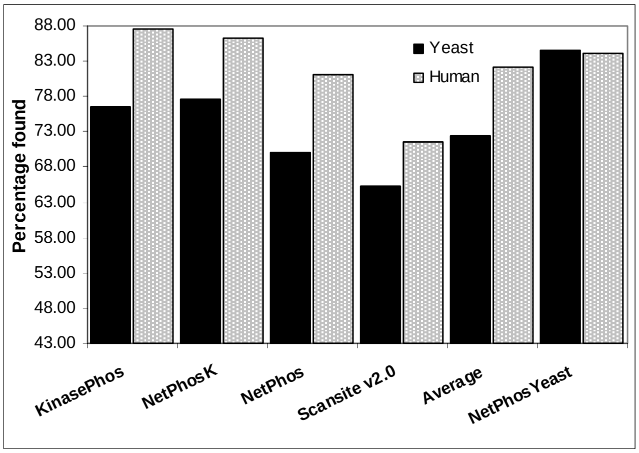

Fig. 1: Sensitivity comparison of KinasePhos, NetPhosK, NetPhos, Scansite v2.0 and NetPhosYeast on the verified human and yeast phosphorylation sites.

The artificial neural network (ANN) used in this study was a standard three-layer feed forward type that has been described previously (Qian et al. 1988). In addition to previous proposed methods for predicting phosphosites amino acids were encoded with the BLOSUM62 scoring matrix (Nielsen et al. 2003) for achieving a more general physicochemical description as compared with the sparse encoding scheme (Qian et al. 1988). In the BLOSUM62 encoding scheme each amino acid is represented by the corresponding vector of numbers in the BLOSUM62 matrix denoting the penalty for replacing that amino acid with any of the 19 others amino acids. The ANN input window size for the sequences and the number of neurons in the hidden layer was subsequently optimized in each cross-validation procedure

3 RESULTS

In total, 20 artificial neural networks were trained to classify validated yeast phosphorylation sites by optimizing input window size and the number of hidden units in each cross-validation set. The average output of these 20 networks constitutes the output score from NetPhosYeast. Using the independent evaluation scheme as described in the methods, NetPhosYeast achieves an average MCC of 0.74, a sensitivity of 0.84 and a specificity of 0.90 using a threshold of 0.5.

To estimate the ability of NetPhosYeast to identify phosphosites in yeast and human we compared its performance with four existing phosphosite predictors, that allow multiple sequence submissions: NetPhos, NetPhosK, KinasePhos and Scansite v2.0 (the respective setting that gives raise to the maximal MCC was used for each prediction method). On average these methods find 82% of all annotated human phosphorylation sites in Swiss-Prot, which is comparable to NetPhosYeast (see Figure 1). This suggests that there is a considerable overlap in recognition sequence space between the kinases repertoire of the two species. Using the independent evaluation data set, NetPhosYeast identifies 84% of yeast phosphosites, whereas the aforementioned methods identify 67% on average. This indicates the existence of yeast-specific substrate motifs, which are exclusively recognized by NetPhosYeast, and demonstrates the need for a yeast-specific predictor.

4 CONCLUSION

The method presented here predicts phosphorylation sites in yeast proteins with high specificity (0.90) and sensitivity (0.84) measured on an independent data set. Since many researchers use yeast as the preferred model organism NetPhosYeast will aid the sequence analysis of proteins in their work. As more data will become available, the next generation of phosphorylation site predictors will move towards both species and kinase specificity

ACKNOWLEDGEMENTS

This project was supported by funds from The Danish National Research Foundation and from EU-FP6: Interaction Proteome.

REFERENCES

Ball, C. A., et al. (2000) Integrating functional genomic information into the Saccharomyces genome database. Nucleic Acids Res, 28, 77-80.

Blom, N., et al. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol, 294, 1351-62.

Blom, N., et al. (1998) PhosphoBase: a database of phosphorylation sites. Nucleic Acids Res, 26, 382-6.

Blom, N., et al. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 4, 1633-49.

Diella, F., et al. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics, 5, 79.

Ficarro, S. B., et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol, 20, 301-5.

Gruhler, A., et al. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol Cell Proteomics, 4, 310-27.

Huang, H. D., et al. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res, 33, W226-9.

Iakoucheva, L. M., et al. (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res, 32, 1037-49.

Jensen, O. N. (2006) Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol, 7, 391-403.

Manning, G., et al. (2002) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci, 27, 514-20.

Obenauer, J. C., et al. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res, 31, 3635-41.

Qian, N., et al. (1988) Predicting the secondary structure of globular proteins using neural network models. J Mol Biol, 202, 865-84.

Sigrist, C. J., et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform, 3, 265-74.

Xue, Y., et al. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics, 7, 163.

Xue, Y., et al. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res, 33, W184-7.