On the Quality of Motifs for Protein Phosphorylation Site Prediction

Yasser EL-Manzalawy [1,2] , Cornelia Caragea [1,2] , Drena Dobbs [3,4,5] , Vasant Honavar [1,2,4,5]

[1] Artificial Intelligence Laboratory, [2] Department of Computer Science
[3] Department of Genetics, Development and Cell Biology
[4] Bioinformatics and Computational Biology Graduate Program
[5] Center for Computational Intelligence, Learning, and Discovery
Iowa State University, Ames, IA 50010, USA.
{yasser, cornelia, ddobbs, honavar}@iastate.edu

Short Abstract

Motif-based bioinformatics tools allow users to set threshold values for specific motifs, even though users may not know how these values affect performance. We propose statistical measures for assessing "motif quality" and the relationship between p-values and true positive rate, using phosphorylation site prediction as a test case.

Long Abstract

Protein phosphorylation is an important post-translational modification that can dramatic alter the biological activity of proteins. Several computational methods for predicting phosphorylation sites for specific protein kinases have been proposed, including motif-based approaches that rely on Position Specific Scoring Matrices (PSSMs) and Hidden Markov Models (HMMs). A PSSM or HMM motif is constructed from an ungapped multiple-sequence alignment that is expected to carry some signal. This motif can be used to score new sequences and the higher the score, the more likely that the new sequence carries the same signal modeled by the motif. In general, each motif has a predetermined threshold score that maximizes the prediction accuracy of the motif on a validation set. However, many motif-based tools allow users to set a different threshold score or to specify a certain false positive rate, p-value, for the motif. For instance, Scansite [1] and KinasePhos [2], two popular tools (that use PSSMs and HMMs respectively) to predict kinase-specific phosphorylation sites, provide several options to modify the threshold score for motifs. A major problem with this approach is that for a chosen p-value, or false positive rate, the user has no way of knowing what the corresponding true positive rate is because the reported performance of the motif corresponds to that obtained using predetermined threshold scores. Against this background, we explore statistical measures for assessing the quality of a motif and the relation between p-values and the true positive rate. These statistical measures are the Receiver Operating Characteristic (ROC) curve and the area under ROC (AUC) which are widely used by machine learning researchers to report the performance of their classifiers.

Because Scansite and KinasePhos motifs are not publicly available to users (except through the online servers that generate predictions based on the motifs), and both methods do not return scores for negative predictions, it is not straightforward to compare the ROC curves for their motifs. Hence, we explored an alternative approach to compare

the two methods. We considered only kinase families with more than 50 reported phoshporylation sites in Phospho.ELM [3]; thus, six kinase families, CDK, CK2, MAPK, PKA, PKB, and PKC, were considered in our experiments. For each family, we extracted positive examples using 15-residue amino acid sequence window centered at known phosphorylated Ser/Thr sites and negative examples using the same window centered at Ser/Thr residues that are not annotated as phosphorylation sites in the *same proteins*. We created a dataset for each family consisting of positive examples for that family and randomly selected negative examples equal to the number of positive examples in that family. We used ProfileWeight [4] to build PSSM motifs and HMMER [5] to build HMM motifs using only the positive examples for each family. We computed the ROC curve and AUC obtained by the 5-fold cross validation: The data set for each kinase family is randomly partitioned into 5 parts of equal size such that the ratio of the positive and negative examples in each part is the same. On each cross validation experiment, we used positive instances in four of the five subsets for building PSSM and HMM motifs and the remaining subset for evaluating the motifs. The reported performance is based on averages across the five cross validation runs. It should be noted that our HMM motifs are different from KinasePhos motifs since KinasePhos uses a window of 9 amino acids and usually builds more than one motif per kinase family by clustering the sequences of large families and building a motif from each cluster.

Our results show that HMM motifs are superior on PSSM motifs for predicting protein phosphorylation sites for the CK2, PKA, and PKC protein kinase families. For the CDK family, both PSSM and HMM motifs have nearly the same AUC, but the HMM has a better true positive rate for p-values ranging from 0% to 6%. In the case of PKB and MAPK, PSSM motifs perform better than HMM motifs. In all cases, visualizing the ROC curve of the motif can assist users in selecting a proper threshold and in interpreting the resulting predictions. Furthermore, the reported quality of the motif based on an evaluation procedure such as the one outlined here can help users in choosing the better performing motif-based prediction tool for a given task.

[1] Obenauer JC, Cantley LC, and Yaffe MB (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31(13): 3635-3641.
[2] Huang HD, Lee TY, Tzeng SW, and Horng JT (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res. 33(1): W226-W229.
[3] Diella F, Cameron S, et al. (2004) Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5(1): 79.
[4] Thompson JD, Higgins DG, and Gibson TJ (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput. Appl. Biosci. 10(1): 19-29.
[5] Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.