Domain-context information for the improvement of phosphorylation site prediction

A. Palmeri✉, M. Helmer-Citterich, P.F. Gherardini

Centre of for Molecular Bioinformatics, Department of Biology, University of Tor Vergata, Roma, Italy

Motivations

Our understanding of the determinants of protein phosphorylation is far from being complete. In many predictive systems, linear-motifs represent the main features that have a high power of discrimination between the phosphosites and the non-phosphorylated sites. The majority of the tools consider structural features like the secondary structure, the disorder or the solvent accessibility of the phosphopeptide. However the protein context in which the phosphopetide is found is in many cases ignored. The aim of this work is to study the distribution of phosphorylation sites with respect to protein domains. Moreover we want to investigate whether including domain information improves the prediction of phosphorylation sites.

Methods

We collected Human Phosphorylation data from the PhosphositePlus [1], phospho.ELM [2], PHOSIDA [3] and Swissprot databases and mapped these phosphosites on the human proteome downloaded from the Swissprot. We identified the domain boundaries on each protein, using the Pfam scanner [4]. We then performed a statistical test of significance for each domain type to identify the domain types that are enriched or depleted in phosphorylation. The significance test outputs the probability that the relative abundance of each domain type is different between the phosphoproteome and the overall proteome. As the majority of phosphosites are located outside protein domains, we performed the same analysis for Inter Domain Regions. We identify a IDR, as the protein region that is enclosed by two specific domains or by one specific domain and the C-terminal or N-terminal of a protein. A phosphosite predictor has been developed using human phosphorylation data. The training and testing procedures were written in R, using the package LiblineaR. The features that we used in the predictor are: the -5/+5 residues around the phosphosite, a Local Domain Feature and a Best Domain Feature. We encoded in the predictor the sequence features in standard orthogonal encoding. The Local Domain Feature represents the propensity of a specific domain type to be phosphorylated. It is calculated from the training data as the proportion of phosphorylated domains of a specific domain type on all the occurrences of that domain type. The Best Domain Feature depends on the whole domain-composition of the protein and it is defined as the number of phosphoproteins in which the domain-type is found divided by the total of the proteins that contain that domain-type. When we predict a site in a protein the Best Domain Feature is represented by the maximum among all the propensities of the domains contained in the protein.

Results

We performed a statistical test to identify domains that are significantly enriched or depleted in phosphorylation. Two tests were performed: one for the Tyr and the other for the Ser/Thr. For the Tyr phosphorylation, we obtained 26 domain types enriched and 19 depleted. For the Ser/Thr phosphorylation we found that there are 22 enriched and 26 depleted domains. The domain types that are enriched and depleted in phosphorylation represent almost the 0.5% of all the existing domain types. But the occurrences of the domains enriched in phosphorylation account for almost the 3% of all the domains in the Human Proteome. Moreover the occurrences of the depleted domains represent almost the 30% of the Human Proteome. However the majority of phosphosites are in protein regions outside domains. Thus, the same analysis was performed on Inter Domain Regions obtaining very similar results. Having observed that domain-context information influences phosphorylation, we wanted to test if it could be useful for phosphosite prediction. Therefore we trained two phosphosite predictors, one using only the sequence and the other with all the domain-contextual features, encoded in the Local Domain Feature and the Best Domain Feature. We compared the performances between these two predictors, using the AUC, in a test dataset independent of the training. The sequence-only predictor obtained an AUC of 0.71 for Ser/Thr prediction and of 0.63 for Tyr prediction. The predictor with all the features reached an AUC of 0.78 for Ser/Thr prediction and of 0.72 for Tyr prediction. In both Ser/Thr and Tyr predictions we observed a 10% improvement in performance, due to the inclusion of the domain-context features.

References

1. Dinkel et al., Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D261-7. Epub 2010 Nov 9.
2. Hornbeck et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse, Nucleic Acids Research, 2011, 1-10 doi:10.1093/nar/gkr1122
3. Gnad et al., Nucleic Acids Res. 2011 Jan;39(Database issue):D253-60. Epub 2010 Nov 16. 4. Finn et al., The Pfam protein families database. Nucleic Acids Res. 2010 Jan;38(Database issue):D211-22. Epub 2009 Nov 17.