Structure based approach for predicting substrates for protein kinases

Debasisa Mohanty
Bioinformatics Center
National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi – 110067
E-mail: deb@nii.res.in

In silico methods are being widely used for identifying substrates for various kinases and deciphering cell signaling networks. However, most of the available phosphorylation site prediction methods use motifs or profiles derived from a known data set of kinase substrates and hence, their applicability is limited to only those kinase families for which experimental substrate data is available. Therefore it is necessary to develop structure based approaches, which do not require training using experimental substrate data.

We have developed a novel structure based approach for predicting substrates of protein kinases (1). The putative substrate peptides are modeled in the substrate binding pockets of kinases using the available crystal structures of kinase-peptide complexes as templates. The binding energy of these peptides in complex with the kinase are evaluated using a residue based statistical pair potential derived by Betancourt and Thirumalai (2). We have carried out detailed benchmarking of this approach on the experimental data available in the Phospho.ELM database (3) and compared our results with those from a number of other phosphorylation site prediction tools. Our results indicate that, the structure based method developed in this work can predict more than 60% of the experimentally identified substrates for 10 protein kinases. The prediction accuracies for PKA, PKB, PKG, and PDK were well above 70% with PKG having the highest prediction accuracy of 81.5%. The other kinase groups for which our approach showed good prediction accuracy were ChK, CK2, DAPK, ROCK, and MAP3K. Our approach also outperformed all other prediction tools for PKG, PDK, ChK, CK2, DAPK, ROCK, and MAP3K. We also carried out receiver operating characteristic (ROC) curve analysis for analyzing the robustness of our structure based prediction approach. The area under curve (AUC) values for these 10 kinases ranged from 0.681 (MAP3K) to 0.838 (PKG). It is encouraging to note that, the prediction accuracy of our method is comparable to other sequence based methods like GPS, PPSP, SCANSITE and NetPhosK, even though it does not use any experimental phosphorylation site data for training unlike sequence based methods.

We have also demonstrated that, percentile score of the true phosphorylation site can be further improved by using a multi scale approach. In this approach, high scoring peptides

short listed by initial pair potential based screening are re-ranked by evaluating their binding energy using all atom MM/PBSA energy function. Interestingly such multi scale approach has also given interesting results in case of MHC-peptide recognition (4) and can in principle be extended to other systems involving peptide recognition modules (PRM) like PTB, PDZ and WW etc.

References

1. Kumar N and Mohanty D (2010) Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials. ***Bioinformatics*** **26**, 189-97.

2. Betancourt, M.R. and Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. ***Protein Sci***, **8**, 361-369.

3. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N. and Gibson, T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. ***BMC Bioinformatics***, **5**, 79.

4. Kumar N and Mohanty D (2010) Structure based identification of MHC binding peptides: Benchmarking of prediction accuracy. **Molecular Biosystems 6**, 2508-2520.