Prediction of protein posttranslational modifications: comparative analysis of known approaches

Boris Sobolev, A.V. Veselovsky, V.V. Poroikov

Orekhovich Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, 119121, Russia Moscow, Pogodinskaya street, 10/8, boris.sobolev@ibmc.msk.ru

Covalent posttranslational modifications (PTM) of proteins are involved into the numerous biological processes, such as cellular signaling, cell adhesion, control of gene expression, protein folding, etc. [1]. Even the modern high-throughput experimental methods do not able to reveal the whole variety of PTMs in living systems. Due to this reason, prediction of PTM *in silico* is widely applied in designing the experiments, advanced studying the complex regulatory processes, simulating the signal transduction networks [2, 3]. Currently, numerous methods are used to recognize the potential PTM sites in amino acid sequences of proteins. We performed a comparative analysis on PTM prediction using 17 freely accessible tools on a case study widely existed PTMs including phosphorylation, acetylation, glycosylation and ubiquitination. Using the representative validation set consisted of 80 diverse proteins we show that utilization of the close surroundings of PTM sites as the only discriminative features does not provide the reasonable accuracy of prediction [3].

The poor accuracy of PTM prediction can be explained by several reasons:

- Modification of the certain PTM is performed by the groups of enzymes different in the substrate specificity. However, the amount of accessible data is not always sufficient to create the representative training sets for different classes of modifying enzymes. Currently, phosphorylation is the best studied type of PTM, which satisfied to this requirement.
- Commonly, one cannot be sure that all functionally significant PTMs are experimentally identified for each studied protein, especially in case of short-term reversible modifications. Some PTM determined *in vitro* may not occur *in vivo*. This uncertainty

results in ambiguous distinction of positive and negative cases.

- Generally, the PTM events are defined by multiple space-time factors regulated the modifications. The short linear modules containing the modified sites determine the specificity to the certain PTM in combination with other regions. These regions are crucial for binding the protein substrate with the modifying enzyme or other protein partners providing the necessary enzyme-substrate interaction. Furthermore, the co-expression and subcellular location of interacting proteins can also define the probability of PTM events.

We come to the conclusion, that the wider sets of data should be implicated for computational prediction of PTM. The application of phylogenic approach seems to be reasonable, if it is possible to collect the representative set of homologues proteins with PTM related to the *same or similar* functional features (e.g., cell regulatory pathways). Usage of data on protein-protein interactions increases the specificity but strongly decreases the sensitivity of prediction. Our analysis demonstrates that more detailed data in the training set provides the basis for increase of the prediction accuracy.

The study was supported by the grant of Russian Ministry of Education and Science (Agreement No. 8274).

1. E. Basle et al. (2010). Protein chemical modification on endogenous amino acids, *Chem. Biol.* **17**:213-227.

2. B. Eisenhaber, F. Eisenhaber (2010). Prediction of posttranslational modification of proteins from their amino acid sequence, In: *Data Mining Techniques for the Life Sciences. Series: Methods Mol. Biol., V. 609.* O.Carugo, F.Eisenhaber (Eds.), 365-384 (Humana Press).

3. B.N. Sobolev et al. (2013). Prediction of posttranslational modifications in proteins: trends and methods, *submitted to Russian Chemical Reviews (Uspekhi Khimii).*