A bioinformatic approach for a structural analysis of protein phosphorylation sites

Zanzoni A, Gherardini F, Ausiello G, Via A, Helmer-Citterich M

Centre for Molecular Bioinformatics, Department of Biology,
University of Rome Tor Vergata.

Motivation

The phosphorylation of specific protein residues is a crucial event in the regulation of several cellular processes, acting on activation, deactivation or recognition of the target protein. A great amount of eukaryotic cell proteins (30 up to 50% of the total) undergoes such post-translational modification. The recent improvement in the experimental identification of phosphoproteins and phosphoresidues has increased dramatically the amount of phosphorylation sites data and the need of computational tools for collecting and analysing this data has grown accordingly. In the past years several sequence-based methods to predict phosphorylation sites were developed using different approaches such as regular expressions with context-based rules, Position-Specific Scoring Matrices (PSSMs) and artificial neural networks. Only approximately one tenth of known kinases have known consensus sequences, which often are not present in all known in vivo substrates. The structural basis and the determinants of interaction specificity are often unclear. The presence of structural determinants that only sometimes overlap with sequence consensi and that might be independent on the residue order in protein sequences might explain the problems encountered so far in unravelling the rules of kinase specificity.

Methods

We have developed a procedure for the annotation and analysis of the three-dimensional structure of experimentally verified protein phosphorylation sites, also called instances, retrieved from the phospho.ELM database. The correspondence between phospho.ELM sequences and the PDB chains was established via the Seq2Struct resource, an exhaustive collection of annotated links between UniprotKB and PDB sequences. Links are based on sequence alignment using pre-established highly reliable thresholds. For each instance mapped onto PDB chains, a structural neighbourhood, that we call zone, was defined using a distance criterion. A procedure was implemented in order to annotate each residue belonging to the defined zones with diverse functional information such as solvent accessibility, secondary structure assignment and sequence conservation. Furthermore, we performed an all versus all comparison amongst the phosphorylation zones in order to find statistically significant local structural similarities.

Results

All this information, as well as the results of a large-scale local structural comparison with stringent parameters, was stored in a publicly available relational database called Phospho3D [http://cbm.bio.uniroma2.it/phospho3d]. Amongst the structural comparison results, we selected two potentially interesting cases: a structural match that allows the inference of which kinase could be responsible for the phosphorylation of a specific tyrosine residue and an interesting candidate 3D motif in common between two substrates. In the latter structural motif, some residues are conserved both in sequence and in structure, some other are only conserved in structure and not in sequence; an experimental test is being designed to evaluate its biological relevance.

Contact email: andreas@cbm.bio.uniroma2.it